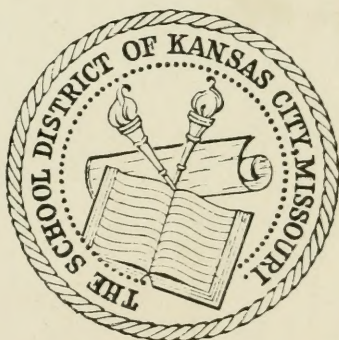


Kansas City
Public Library



This Volume is for
REFERENCE USE ONLY

PUBLIC LIBRARY
KANSAS CITY
MO

From the collection of the

j f d
y z n m k
x
o PreLinger a h
u v q g Library
e
b t s w p c

San Francisco, California
2008

YBARRIL 31884
YTO 2A2843
084

PUBLIC LIBRARY
KANSAS CITY
MO

YHABU 3100H
YTD 3400H
OH

KAN-
FEB 3 - 1953
THE BELL SYSTEM
TECHNICAL JOURNAL

A JOURNAL DEVOTED TO THE
SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL
COMMUNICATION

ADVISORY BOARD

S. BRACKEN

F. R. KAPPEL

M. J. KELLY

EDITORIAL COMMITTEE

E. I. GREEN, *Chairman*

A. J. BUSCH

F. R. LACK

W. H. DOHERTY

J. W. McRAE

G. D. EDWARDS

W. H. NUNN

J. B. FISK

H. I. ROMNES

R. K. HONAMAN

H. V. SCHMIDT

EDITORIAL STAFF

PHILIP C. JONES, *Editor*

M. E. STRIEBY, *Managing Editor*

R. L. SHEPHERD, *Production Editor*

INDEX

VOLUME XXXI

1952

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

V. 31: Jan - Nov. 52
H E B E L L S Y S T E M

Technical Journal

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

VOLUME XXXI

JANUARY 1952

NUMBER 1

The Ferromagnetic Faraday Effect at Microwave Frequencies and
its Applications—The Microwave Gyrator C. L. HOGAN 1

Dialing Habits of Telephone Customers
CHARLES CLOS AND ROGER I. WILKINSON 32

Selective Fading of Microwaves
A. B. CRAWFORD AND W. C. JAKES, JR. 68

Propagation Studies at Microwave Frequencies by Means of Very
Short Pulses O. E. DE LANGE 91

Properties of Ionic Bombarded Silicon RUSSELL S. OHL 104

Mechanical Properties of Polymers at Ultrasonic Frequencies
WARREN P. MASON AND H. J. MCSKIMIN 122

Relay Armature Rebound Analysis ERIC EDEN SUMNER 172

Abstracts of Bell System Technical Papers Not Published in This
Journal 201

Contributors to This Issue 213

THE BELL SYSTEM TECHNICAL JOURNAL

PUBLISHED SIX TIMES A YEAR BY THE
AMERICAN TELEPHONE AND TELEGRAPH COMPANY

195 BROADWAY, NEW YORK 7, N. Y.

CLEO F. CRAIG, *President*

CARROLL O. BICKELHAUPT, *Secretary*

DONALD R. BELCHER, *Treasurer*

EDITORIAL BOARD

F. R. KAPPEL

O. E. BUCKLEY

H. S. OSBORNE

M. J. KELLY

J. J. PILLIOD

A. B. CLARK

R. BOWN

D. A. QUARLES

F. J. FEELY

PHILIP C. JONES, *Editor*

M. E. STRIEBY, *Managing Editor*

SUBSCRIPTIONS

Subscriptions are accepted at \$3.00 per year. Single copies are 75 cents each.

The foreign postage is 65 cents per year or 11 cents per copy.

PRINTED IN U.S.A.

THE BELL SYSTEM TECHNICAL JOURNAL

VOLUME XXXI

JANUARY 1952

NUMBER 1

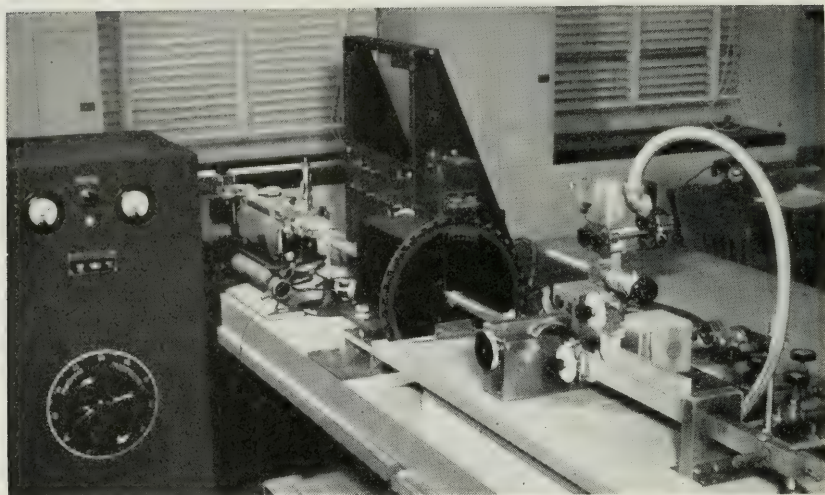
The Ferromagnetic Faraday Effect at Microwave Frequencies and its Applications

The Microwave Gyrotator

BY C. L. HOGAN

A new microwave circuit element dependent on the Faraday rotation of a polarized wave has been developed. The element violates the reciprocity theorem and, because it shares this property with a gyroscope and because it is dependent on gyromagnetic resonance absorption, it has been termed a microwave gyrotator. It is a low-loss broadband device with many applications. Among these are one-way transmission systems, microwave circulators, microwave switches, electrically controlled variable attenuators and modulators.

The microwave gyrotator has been realized by making use of the Faraday rotation in pieces of ferrite placed in the waveguide. Polder has previously shown, in his analysis of the gyromagnetic resonance phenomenon, that ferromagnetic substances should show appreciable Faraday rotations at microwave frequencies. In the present study, Polder's analysis has been extended to include a wave being propagated through a ferromagnetic substance with dielectric and magnetic loss, and data are presented which give experimental verification of the theory. In addition an experimental technique is described which may be of some interest in studying the properties of ferrites at microwave frequencies.



Photograph of the experimental setup shown diagrammatically in Fig. 5.

INTRODUCTION

In a recent series of articles, Tellegen¹ has discussed the possible applications of a new circuit element which he calls a gyrator. He defines the ideal gyrator, in principle, as a passive four-pole element which is described by: (see Fig. 1)

$$v_1 = -Si_2 \quad v_2 = Si_1 \quad (1)$$

Since the coefficients above are of opposite sign, the gyrator violates the theorem of reciprocity. Any network composed of the usual electrical circuit elements—resistors, inductors, capacitors, and transformers—will satisfy the theorem of reciprocity. In simple terms, this theorem states that if one inserts a voltage at one point in the network and measures the current at some other point, their ratio (called the transfer impedance) will be the same if the positions of voltage and current are interchanged. In the gyrator, however, this transfer impedance for one

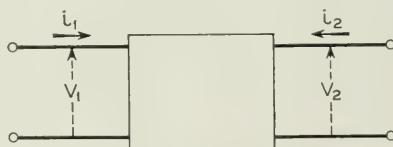


Fig. 1—General four-pole.

direction of propagation is the negative of that for the other direction of propagation. Essentially this means that a 180° phase difference exists between the two directions of propagation. For this reason it has been suggested that the element could be more aptly called a directional inverter.²

Network synthesis today is based upon the existence of four basic circuit elements: the capacitor, the resistor, the inductor, and the ideal transformer. It is apparent that the introduction of a fifth circuit element, the gyrator, would lead to considerably improved solutions for many network problems. In fact, Tellegen¹ has shown that the synthesis of resistanceless four-pole networks would be much simplified by its introduction. In addition, McMillan³ has shown that it would be possible to construct a one-way transmission system if a gyrator were available, and Miles⁴ has shown that it would be possible by use of a gyrator to construct a network which is equivalent to a Class A vacuum tube amplifier circuit. While the realizable power gain of these gyrator circuits is necessarily always not greater than unity, many other networks including gyrators are possible which have properties analogous to vacuum tube circuits and some of these may be of practical importance. Since this new element offers such interesting possibilities in network synthesis, a study has recently been made in these Laboratories of possible methods for realizing the gyrator.

A gyrator was employed by Bloch⁵ in his measurement of the magnetic moment of the proton. Bloch made use of the phenomenon that if two crossed coils with a mutual core are adjusted so that there is zero mutual inductance between them and if a steady magnetic field is applied perpendicular to the axes of both coils, then an ac voltage applied to one of the coils will induce a voltage in the second due to the gyromagnetic resonance phenomenon. This induction is ordinarily extremely small unless the magnetic field is adjusted so that the exciting frequency coincides with a gyromagnetic resonance frequency of the material which forms the mutual core of the two coils. In Bloch's experiment, the magnetic field was held constant and the exciting frequency was adjusted until it coincided with the gyromagnetic resonance frequency of the proton. If they were wound over a paramagnetic or ferromagnetic material, the two crossed coils would form a gyrator when the magnetic field was adjusted so that the frequency of the exciting field coincided with the gyromagnetic resonance of the unpaired electrons. The fact that this structure constituted a gyrator was first recognized by Tellegen¹ and has been discussed by Beljers and Snoek⁶

in a paper which gives a very satisfying physical model with which to interpret gyromagnetic phenomena occurring within ferrites.

Physical analysis indicates that the properties of ferromagnetic materials can be explained by assuming that the electron behaves as if it were a negatively charged sphere which is spinning about its own axis with a fixed angular momentum. This rotation of charge imparts to the electron a magnetic moment which is a function of the electric charge on the electron, the angular velocity of the electron, and its size. Thus the electron behaves as if it were a spinning magnetic top, whose magnetic moment lies along the axis of rotation, and its behavior can be understood by considering a spinning gyroscope suspended in gimbal rings at a point not coinciding with its center of gravity. If a gyroscope, thus supported in a gravitational field, is lifted away from its position of minimum potential energy and then released, it will not return to the position of minimum energy but will precess about the vertical axis. This is illustrated in Fig. 2 where the spinning gyroscope makes an angle θ with the vertical z , axis. Its equilibrium motion, in the absence of damping, is a precessional motion about the vertical axis with a velocity ω_p .

If the gyroscope be regarded as initially hanging vertically downward

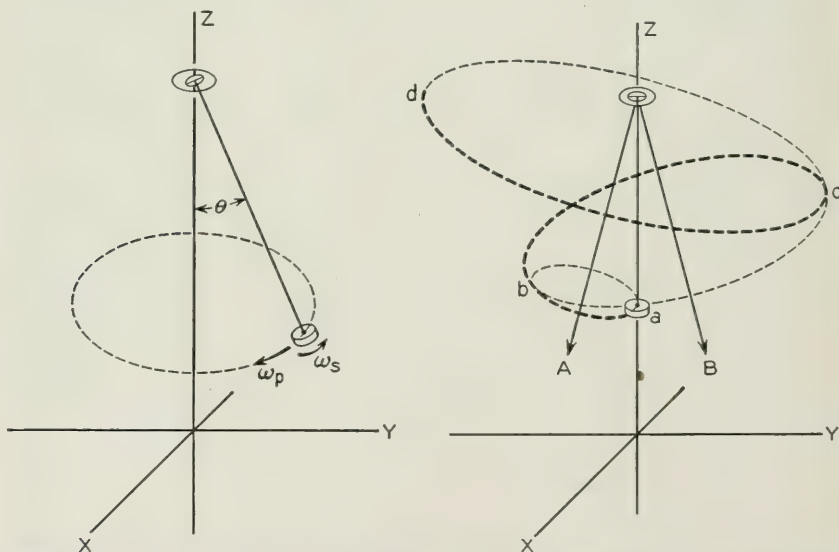


Fig. 2 (left)—Precessional motion of a gyroscopic pendulum in a gravitational field. Fig. 3 (right)—Precessional motion of a gyroscopic pendulum in a gravitational field which oscillates between the directions A and B.

as indicated in Fig. 3 and then a gravitational force is suddenly made to act along the y axis so that the net gravitational force acts along A , it is obvious that the gyroscope will begin to precess about the gravitational field direction as indicated by the small dotted circle. However, if after completing a half cycle, the horizontal component of the gravitational field is reversed so that now the net gravitational field acts along the vector B , the gyroscope will begin to precess about B as indicated by the intermediate size dotted circle. If the horizontal component of the gravitational field is again reversed after the gyroscope completes another half cycle in its precession, the gyroscope will again begin to precess about the direction A and the actual path of the precessional motion will be along the path $a-b-c-d$. If this process is continued indefinitely, the gyroscope will precess in larger and larger circles around the vertical until the damping becomes large enough to contain the gyroscope in some equilibrium circle (assuming that the damping is large enough to accomplish this).

The above model affords a classical picture which can be used quite readily to describe the motion of the electrons in a ferrite. If the ferrite is initially saturated along the z axis by a steady magnetic field, the electrons will come to rest with their magnetic moments lying along the z axis, as the gyroscope in Fig. 3. If now an alternating magnetic field is applied along the y axis, the electrons will begin to precess in larger and larger circles about the z axis until they finally reach some equilibrium position under the influence of the magnetic fields and the damping. Thus it is apparent in the gyromagnetic resonance experiments described above why an alternating field applied perpendicular to a steady magnetic field in a ferrite will give rise to a varying flux perpendicular to both the steady field and the alternating field. It is also apparent why the alternating flux along the x axis is 90° out of phase with the alternating flux along the y axis. Since precession of the top will always be in the same direction regardless of whether the alternating field is applied along the x or y axes, consideration of Fig. 3 makes it apparent how the two crossed coils with ferrite at their center can constitute a gyrator which violates the reciprocity relation in a manner described by Equations (1). To the present time, however, no practical circuit element making use of this phenomenon has been constructed because the coefficient of coupling between the coils is always small, even in the vicinity of the resonant frequency, and also because the losses in the materials available are so high in the vicinity

of resonance that the insertion loss of such a device would be prohibitively large.

McMillan in his original article,³ showed that a gyrator could be realized by means of mechanically coupled piezo-electric and electromagnetic transducers. Later, McMillan⁷ pointed out that a gyrator could be realized by means of the Hall effect in a square plate of bismuth, as was also predicted by Casimir.⁷ Another similar possibility would be an electrical-electrical coupling through a gyroscopic link. A gyrator has been built by W. P. Mason of these Laboratories which makes use of the Hall effect in a crystal of germanium.⁸ This gyrator showed an insertion loss somewhat higher than the theoretical loss of 12.3 db. R. O. Grisdale of these Laboratories suggested that these losses could be greatly reduced if the same Hall effect principle were applied to a vacuum tube which contained four electrodes which could both emit and collect electrons. This device is no longer passive, but such a structure has been built and showed an insertion loss of about 7 db, only slightly higher than the theoretical loss which would be expected from this geometry.

In view of the substantial losses found to exist in the earlier forms of gyrator discussed above, a study of other "anti-reciprocal" phenomena which might lead to the realization of a relatively low loss gyrator was undertaken.

It has long been known that the Faraday rotation of the plane of polarization in optics is anti-reciprocal. In order to observe the Faraday rotation, polarized electromagnetic waves must be transmitted through a transparent isotropic medium parallel to the direction of the lines of force of a magnetic field. The effect is usually produced by placing the material along the axis of a solenoid. The rotation is "positive" if it is in the direction of the positive electric current which produces the field and "negative" if in the opposite direction. All optically transparent substances show the Faraday rotation.

Its anti-reciprocal property distinguishes the Faraday effect from optical rotations caused by birefringent crystals, or by the Cotton-Mouton effect, which are reciprocal. That is, if a plane polarized light-wave is incident upon a birefringent crystal in such a manner that the plane of polarization is rotated through an angle θ in passing through the crystal, then this rotation will be cancelled if the wave is reflected back through the crystal to its source. In the Faraday rotation, however, the angle of rotation is doubled if the wave is reflected back along its path. Hence, if the length of path through the "active" material is adjusted so as to give a 90° original rotation, the beam on being reflected

will have its plane of polarization rotated a total of 180° in passing in both directions through the material. Thus, the Faraday rotation in optics affords an anti-reciprocal relation quite analogous to the anti-reciprocal property of the gyrator postulated by Tellegen.

Lord Rayleigh⁹ described a one-way transmission system in optics which makes use of the Faraday rotation. Lord Rayleigh's "one-way" system consisted of two polarizing Nicol prisms (oriented so that their planes of acceptance made an angle of 45° with each other), with the material causing the Faraday rotation placed between them. Thus, light which was passed by the first crystal and whose plane of polarization was rotated 45° would be passed by the second crystal also. But, in the reverse direction, the rotation would be in such a sense that light which was admitted to the system by the second crystal would not be passed by the first.

Although Rayleigh's one-way transmission system can be actually realized, it is experimentally difficult since most substances show extremely small Faraday rotations. In fact, large rotations for transparent substances in the optical region are of the order of one degree per cm path length for an applied magnetic field of 1000 oersteds. To realize a rotation of 45° would require maintaining a field of 1000 oersteds over a distance of approximately one-half meter. The Faraday effect in ferromagnetic substances, however, is unique in that it shows rotations many orders of magnitude greater than the rotations exhibited by any other substances. For instance, König¹⁰ reports rotations of $382,000^\circ/\text{cm}$ by passing light through thin layers of magnetized iron. These data, of necessity, however, were taken on extremely thin sections and the total rotation obtained for any specimen did not exceed 10° . In order to obtain appreciable rotations in a device of practical size, it is necessary to obtain a material which shows a rotation at least intermediate between those reported for iron and other ordinary materials. In addition, in order to make effective use of these rotations, the material must be transparent to the radiation which is being used.

THEORY OF THE FERROMAGNETIC FARADAY EFFECT

Polder¹¹ has shown in his analysis of the ferromagnetic resonance phenomenon, that a plane electromagnetic wave at microwave frequencies should show appreciable Faraday rotation when propagated through a ferromagnetic material which is magnetized in a direction parallel to the direction of propagation of the wave. Polder has neglected both magnetic and dielectric losses in his analysis and although for the ferrites which are of greatest interest, this approximation is quite

valid, nevertheless the more complete theory is developed below. The exact theory of this phenomenon should, of course, be approached through quantum mechanics, but since the classical theory, in this particular case, gives a result as satisfactory as the quantum theory and since it lends itself more aptly to a fundamental physical interpretation of the phenomenon, it is the classical theory which is developed here. Quantum mechanically, Faraday rotations in the optical region are accounted for by the Zeeman splitting of the spectral lines.

The classical model which proves quite adequate for the description of ferromagnetic resonance is that illustrated in Figs. 2 and 3 which regards the electrons of the material which contribute to the magnetism as being spinning magnetic tops. The angular momentum of each electron is:

$$|J| = \frac{1}{2}(h/2\pi) \quad (2)$$

\vec{J} = Angular momentum of electron (gm cm²/sec)

h = Planck's constant (6.62×10^{-27} erg sec)

The magnetic moment which arises due to this rotation is:

$$|\mu_B| = \frac{eh}{4\pi mc} \quad (3)$$

where:

$\vec{\mu}_B$ = Magnetic moment of electron (Bohr magneton)

e = Charge on electron (4.80×10^{-10} E.S.U.)

m = Mass of electron (9.10×10^{-28} gm)

c = Velocity of light (3×10^{10} cm/sec)

The so-called gyromagnetic ratio of the electron is the ratio of these quantities and is given by:

$$\gamma = 2 \frac{e}{2mc} = \frac{|\mu_B|}{|J|} \quad (4)$$

If a steady magnetic field is applied to the sample such that the electron sees an effective field H , then a torque will be applied to the electron which tries to turn the electron so that its magnetic moment lies along the field direction. However, as indicated in Fig. 2, the electron will precess around the field direction until damping forces dissipate the energy of precession. The equation of motion of the electron is:

$$\vec{\mu}_B \times \vec{H} = \frac{d\vec{J}}{dt} = \gamma^{-1} \frac{d\vec{\mu}_B}{dt} \quad (5)$$

The equation of motion of the magnetization per unit volume can thus be written:

$$\frac{d\vec{M}}{dt} = \gamma \vec{M} \times \vec{H} \quad (6)$$

where:

\vec{M} = Magnetization of medium

\vec{H} = Macroscopic internal magnetic field

The above equation, however, does not include damping. The damping force, regardless of its origin, must be so introduced into the above equation that it tends to cause the electron's axis of rotation to line up with the field direction. It has been shown by Yager, Galt, Merritt and Wood¹² that the shape of the resonance absorption line can be accounted for if the damping term is introduced in the following way:

$$\frac{d\vec{M}}{dt} = \gamma \vec{M} \times \vec{H} - \frac{\gamma\alpha}{|\vec{M}|} [\vec{M} \times (\vec{M} \times \vec{H})] \quad (7)$$

The vector $\vec{M} \times (\vec{M} \times \vec{H})$ is simply a vector which is in the proper direction to act as a damping force (torque) and the coefficient is chosen so as to give the correct units along with the parameter, α , which must be determined experimentally and which gives the magnitude of the damping torque.

Equation 7 then is the equation of motion of the magnetization of an arbitrarily shaped body under the action of an arbitrary internal field, H . In the appendix, it is shown that if a steady magnetic field, H_a , is applied along the z axis and then a small alternating field is applied in an arbitrary direction to a sample which is infinite in size, the equation relating the resulting alternating flux density, b , and the applied alternating field, h is:

$$\begin{aligned} b_x &= \mu h_x - jKh_y \\ b_y &= jKh_x + \mu h_y \\ b_z &= h_z \end{aligned} \quad (8)$$

where

$$\mu = \mu' - j\mu'' \quad (9)$$

$$K = K' - jK'' \quad (10)$$

Equations which give μ and K in terms of the applied magnetic field and fundamental atomic constants are given in the appendix.

Equations (8) are easily interpreted in terms of the spinning gyroscope model of Fig. 3. If magnetic losses had been ignored (i.e. $\alpha = 0$) then both μ and K would have been real. Under this condition, it is seen that if an alternating field, h_y , is applied along the y axis, then an alternating flux, b_y , is created along the y axis which is in phase with h_y , and an alternating flux, b_x , is created which is 90° out of phase with h_y . Reciprocity between the x and y directions would demand that both terms containing jK should have the same sign. Thus, Equations (8) give a quantitative expression for the results which were previously qualitatively deduced by means of the electronic model illustrated in Figs. 2 and 3.

If a waveguide is filled with a ferromagnetic material such as a ferrite and if then a steady magnetic field is applied along the axis of the waveguide, it is necessary in order to describe this wave to find a solution to Maxwell's equations which is consistent with Equations (8) and in which b , h , E and D are all proportional to $\exp[j\omega t - \Gamma z]$. This problem is not solved exactly. However, in the appendix a solution is obtained for an infinite plane wave. It is found that the ferromagnetic medium can support only a positive or a negative* circularly polarized wave or a combination of both. It is also shown in the appendix that the propagation constants for these two circularly polarized waves are different and are given by the following expressions:

$$\Gamma_+ = \frac{j\omega}{c} \sqrt{(\mu + K)[\epsilon]} \quad (11)$$

and

$$\Gamma_- = \frac{j\omega}{c} \sqrt{(\mu - K)[\epsilon]} \quad (12)$$

where

Γ_{\pm} = Propagation constant

ω = Angular frequency of wave

c = Velocity of light in unbounded space (3×10^{10} cm/sec)

ϵ = Complex dielectric constant of medium

In Equations (11) and (12) it is apparent that the effective permeability of the medium to a positive circularly polarized wave, for in-

* The usual notation is used here, where the positive component is the component which is rotating in the direction of the positive electric current which creates the steady longitudinal field.

stance, is given by the expression $(\mu + K)$, and not by the usual permeability, $b_x/h_x = \mu$. It is also apparent that the quantity $\mu + K$ can vary over wide limits in the vicinity of the ferromagnetic resonance. For this reason, care must be taken in interpreting permeability data for ferromagnetic materials which now occur in the literature and which were obtained by means of impedance measurements at microwave frequencies, since the above equations indicate that this method does not measure the same quantity that is measured at low frequencies by means of a toroidal sample overwound with two coils. The low frequency measurement of permeability obviously measures the quantity which is designated as μ in Equation (8).

If Equations (11) and (12) are solved for the attenuation constants, α_{\pm} , and the phase constants, β_{\pm} , the following results are obtained:

$$\alpha_{\pm} = \frac{\omega}{c} \sqrt{\frac{(\mu' \pm K')\epsilon'}{2}} \cdot \left[\left\{ \sqrt{(1 + \tan \delta_m [4 \tan \delta_d + \tan \delta_m (1 + \tan^2 \delta_d)] + \tan^2 \delta_d)} \right. \right. \\ \left. \left. \cdot - 1 - \tan \delta_m \tan \delta_d \right\}^{\frac{1}{2}} \right] \quad (13)$$

and

$$\beta_{\pm} = \frac{\omega}{c} \sqrt{\frac{(\mu' \pm K')\epsilon'}{2}} \cdot \left[\left\{ \sqrt{(1 + \tan \delta_m [4 \tan \delta_d + \tan \delta_m (1 + \tan^2 \delta_d)] + \tan^2 \delta_d)} \right. \right. \\ \left. \left. \cdot + 1 + \tan \delta_m \tan \delta_d \right\}^{\frac{1}{2}} \right] \quad (14)$$

where:

$$\tan \delta_m = \frac{\mu'' \pm K''}{\mu' \pm K'}$$

(The + sign must be used for a positive circularly polarized wave; the negative sign for the negative circularly polarized wave.)

$$\tan \delta_d = \frac{\epsilon''}{\epsilon'} = \text{dielectric loss tangent}$$

$$\epsilon = \epsilon' - j\epsilon'' = \text{complex dielectric constant}$$

It is almost impossible to get a feeling for what these equations mean with respect to a wave travelling through the medium, especially since μ and K are given by equations which are almost as difficult to perceive. An appreciation of these equations can be obtained however, by reference to Fig. 4 which gives qualitatively the behavior predicted by these expressions. Essentially, α and β are functions of two variables. These are ω , the frequency of the wave, and H_a , the applied magnetic field. In Fig. 4, the index of refraction and attenuation of the positive circularly polarized component are given relative to these values for

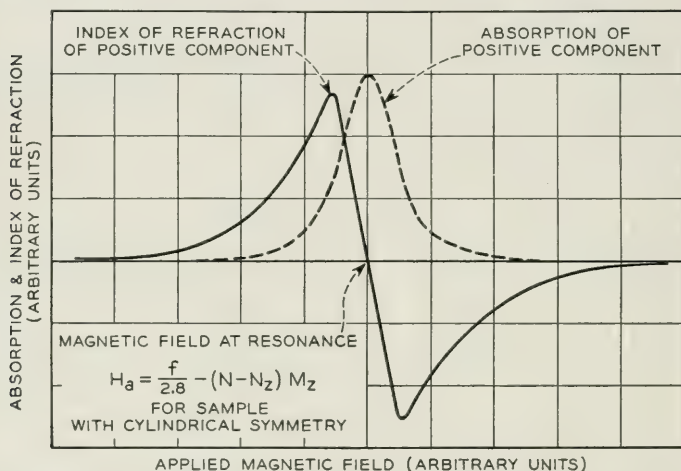


Fig. 4—Index of refraction and absorption of a positive circularly polarized wave relative to the same quantities for a negative circularly polarized wave being propagated through a magnetized medium.

the negative component. Hence, both the index of refraction and attenuation of the negative component are represented by the abscissa of the graph. In Fig. 4 these quantities are plotted as a function of the applied magnetic field for a wave of a fixed frequency. Many of the properties of the medium are clearly displayed in this graph. In particular, as the field necessary for ferromagnetic resonance is approached, the attenuation of the positive component becomes larger and larger. Eventually this component will be substantially completely absorbed and only the negative circularly polarized component will be propagated. Hence it should be possible to establish a circularly polarized wave in a waveguide simply by passing the dominant mode through a ferromagnetic material which is subjected to a longitudinal magnetic field of the proper amplitude. However, there will be an absorption of one-half of the power being propagated. If Fig. 4 had been plotted as a function

of the frequency of the wave for a fixed magnetic field, a similar set of curves would have resulted. This set would indicate the frequency dependence of the Faraday rotation. If the frequency of the wave is far removed from the resonance frequency, the difference between the indices of refraction of the positive and negative component is not frequency dependent. However, near resonance, this difference is a very rapidly varying function of the frequency. It is to be remembered that these equations were derived for an infinite plane wave. However, it would be expected that these equations would describe quite accurately the propagation of the dominant mode in a waveguide. The approximation would, of course, be better when the cut-off wavelength was much greater than the unbounded wavelength. This condition is met when the waveguide is filled with ferrite and for these cases quantitative agreement is obtained.

The above analysis shows that if a dominant mode wave (plane polarized) is incident upon a ferromagnetic material which is magnetized along the length of the waveguide, the wave will split into positive and negative circularly polarized waves whose phase constants are given by Equation (14). Since the two circular components travel with different velocities in the medium, they will upon emerging from it unite to form a plane polarized wave whose plane of polarization has been rotated with respect to the incident polarization. The angle of rotation of the polarization is given by:

$$\theta = \frac{\ell}{2} [\beta_- - \beta_+] \quad (15)$$

where:

ℓ = path length through ferromagnetic material (cm)

In order to evaluate Equation (15), it must be combined with Equation (14). However, a few approximations are valid in Equation (14) which make it much simpler. In particular many ferrites exist for which the magnetic losses are extremely small as long as the internal field within the body is kept small so that the frequency of the wave does not approach the ferromagnetic resonance frequency. This field can be kept small if the magnetic field is not raised above the point necessary to saturate the ferrite. Kittel has shown that for a finite body the effective internal magnetic field that determines the resonant frequency is given by:

$$H_e^2 = [H_a + (N_x - N_z)M_z][H_a + (N_y - N_z)M_z]$$

where H_e is in oersteds. The ferromagnetic resonance frequency is given by:

$$f_{\text{res}} = 2.8H_e \sqrt{1 + \alpha^2} \text{ megacycles} \quad (16)$$

It is easily shown that the following formula is approximately valid for a ferromagnetic body with a circular or square cross-section

$$H_e = \frac{4\pi + N(\mu - 1)}{4\pi + N_z(\mu - 1)} H_a \quad (17)$$

at saturation. Where:

μ = true dc permeability at saturation

N = demagnetizing factor in x and y directions.

If an average value of 1000 is assumed for the dc permeability, then H_e can be readily computed for various shapes.

For a thin disc:

$$N = 0 \quad N_z = 4\pi$$

and

$$H_e = \frac{H_a}{1000}$$

If a thin disc saturates at 1500 gauss, then:

$$H_e = 1.5 \text{ oersteds}$$

$$f_{\text{res}} \approx 4.2 \text{ megacycles}$$

For a long thin pencil:

$$N_z = 0 \quad N = 4\pi$$

and

$$H_e = 1000H_a$$

For this case the body could be saturated with a field of about 1.5 oersteds, so:

$$H_e = 1500 \text{ oersteds}$$

and

$$f_{\text{res}} = 4200 \text{ megacycles}$$

If, for this case, the resonance frequency is so close to the operating frequency that losses due to ferromagnetic resonance become pro-

hibitive, it is wise to then raise the applied field to some high value, so that the resonance frequency will fall well above the operating frequency. Thus, for many cases of interest it is possible by various means to place the ferromagnetic resonance absorption frequency sufficiently far from the operating frequency so that magnetic losses due to this phenomenon are negligible.* The data accumulated to date indicate that the major component of the magnetic losses at microwave frequencies is due to this phenomenon. Only in a few cases have data been taken which have indicated that other factors, such as domain wall relaxation, contribute to the magnetic loss at microwave frequencies.

If then, the magnetic field is controlled so that the ferromagnetic resonance absorption is negligible, Equations (13) and (14) can be simplified to:

$$\alpha_{\pm} = \frac{\omega}{c} \sqrt{\frac{(\mu' \pm K')\epsilon'}{2}} \sqrt{\sqrt{1 + \tan^2 \delta_d} - 1} \quad (18)$$

and:

$$\beta_{\pm} = \frac{\omega}{c} \sqrt{\frac{(\mu' \pm K')\epsilon'}{2}} \sqrt{\sqrt{1 + \tan^2 \delta_d} + 1} \quad (19)$$

which can be written as:

$$\alpha_{\pm} = \frac{\omega}{c} \sqrt{\frac{|\epsilon| - \epsilon'}{2}} \sqrt{\mu' \pm K'} \quad (20)$$

and

$$\beta_{\pm} = \frac{\omega}{c} \sqrt{\frac{|\epsilon| + \epsilon'}{2}} \sqrt{\mu' \pm K'} \quad (21)$$

where μ' and K' are given in the appendix.

If Equation (21) is now inserted into Equation (15) a formula for rotation is obtained which is valid within the limits of the above approximations. If in addition, the frequency of the wave is sufficiently greater than the resonance frequency, so that:

$$\omega_{res} \ll \omega \quad (22)$$

then Equation (15) takes the particularly simple form:

$$\frac{\theta}{l} = \frac{\omega}{2c} \sqrt{\frac{|\epsilon| + \epsilon'}{2}} \left[\sqrt{1 + \frac{4\pi M_z \gamma}{\omega}} - \sqrt{1 - \frac{4\pi M_z \gamma}{\omega}} \right] \quad (23)$$

Most ferrites saturate at 2,000 gauss or less. Hence, for a frequency of

* This is not always possible, for some ferrites, in the polycrystalline state, exhibit extremely broad ferromagnetic resonance absorption lines and it is difficult to operate at any frequency without appreciable absorption.

9,000 megacycles,

$$\frac{4\pi M_z \gamma}{\omega} \leq \frac{2000 \times 17.6 \times 10^6}{9000 \times 2\pi \times 10^6} = 0.622 \quad (24)$$

Hence, the following approximation will be valid to within 5 per cent.

$$\sqrt{1 \pm \frac{4\pi M_z \gamma}{\omega}} \approx 1 \pm \frac{1}{2} \left(\frac{4\pi M_z \gamma}{\omega} \right)$$

With this approximation, Equation (23) reduces to:

$$\frac{\theta}{\ell} = \frac{1}{2c} \sqrt{\frac{|\epsilon| + \epsilon'}{2}} [4\pi M_z \gamma] \quad (25)$$

Equation (25) is quite remarkable. Not only does it predict large rotations, but it also predicts that, within the above approximations the rotation will not depend upon the frequency of the incident radiation. For the assumed values,

$$\epsilon' = 15$$

$$\epsilon'' = 0$$

$$4\pi M_z = 1000,$$

Equation (25) predicts rotations of,

$$\frac{\theta}{\ell} = 65^\circ/\text{cm.}$$

DESCRIPTION OF EQUIPMENT AND MEASURING TECHNIQUES

The Faraday rotation has been measured in a large number of ferrites in order to verify the above theory and in an effort to improve the characteristics of the microwave gyrator. A diagram of the experimental equipment is given in Fig. 5, and a diagram of the test chamber in which the rotations were measured is given in Fig. 6. In the test chamber, two rectangular waveguides are separated by a circular waveguide, the proper nonreflective transitions being made at each end of the circular section, which is about twelve inches long. One rectangular guide is supported so that it can be rotated about the longitudinal axis of the system. The dominant TE_{10} mode is excited in one rectangular guide, and by means of the smooth transition this goes over into the dominant TE_{11} mode in the circular guide. The rectangular guide on the opposite end will accept only that component of the polarization which coincides with the TE_{10} mode in that guide, the other component being reflected

at the transition. Absorbing vanes, inserted in the circular section, absorb this reflected component. The circular guide is placed in a solenoid to establish an axial magnetic field along its length.

The ferrite cylinders to be measured were placed at the mid-section of the circular guide. When a cylinder was used which did not fill the cross-section of the guide, it was supported along the axis of the guide by means of a hollow polystyrene cylinder which did fill the guide.

In addition to measuring the Faraday rotation, measurements of insertion loss were made by determining the power transmitted under identical conditions with the ferrite cylinder removed, and the ellipticity of the transmitted wave was determined by measuring the power transmitted when the rectangular guide on the detector side was rotated to both positions of maximum and minimum transmission. Power transmission measurements could be repeated within 0.2 db. Measurements of the angle of rotation of the plane of polarization could be repeated within $\frac{1}{2}^\circ$ except in the region close to the gyromagnetic resonance where rotations were large and ellipticity so great that it was difficult to decide the positions of maximum and minimum transmission. These errors increased up to the point where the transmitted wave was circularly polarized where it was impossible to measure the angle of rotation.

EXPERIMENTAL RESULTS

Equation (25) indicates that the rotation per unit path length through the ferromagnetic material is proportional to the magnetization of the

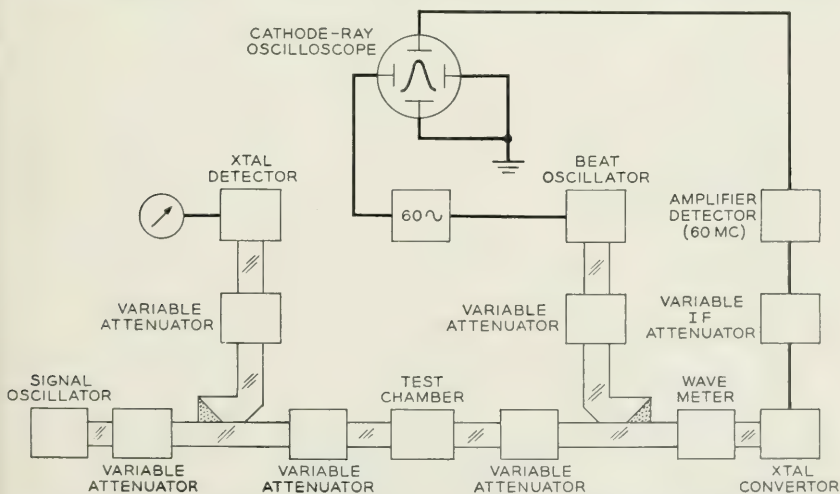


Fig. 5—Experimental equipment set-up used to measure Faraday rotations.

sample and is not dependent directly on the applied magnetic field. Fig. 7 shows the dependence of rotation upon magnetization for a sample of manganese zinc ferrite, and indicates that after the sample is saturated, the rotation is sensibly independent of the applied magnetic field. In addition, the complex dielectric constant and the saturation magnetization of this sample were measured. From these the rotation per centimeter path can be computed from the above theory using

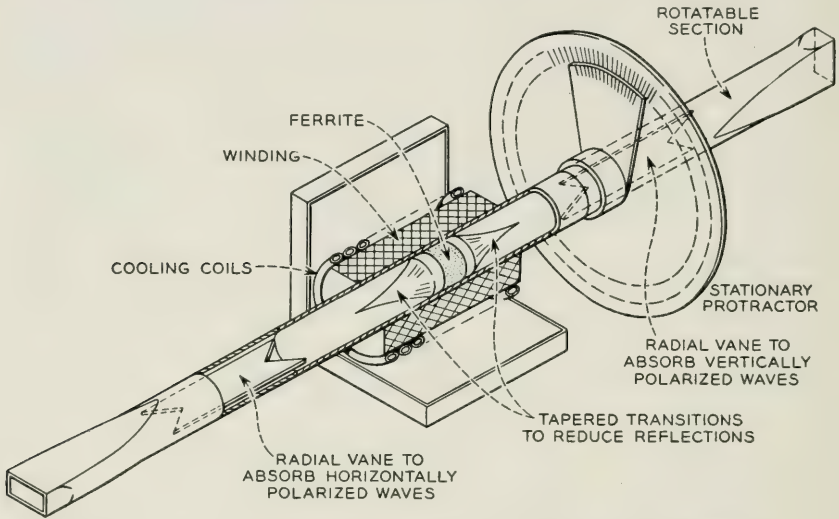


Fig. 6—Detail of test chamber in which rotations were measured.

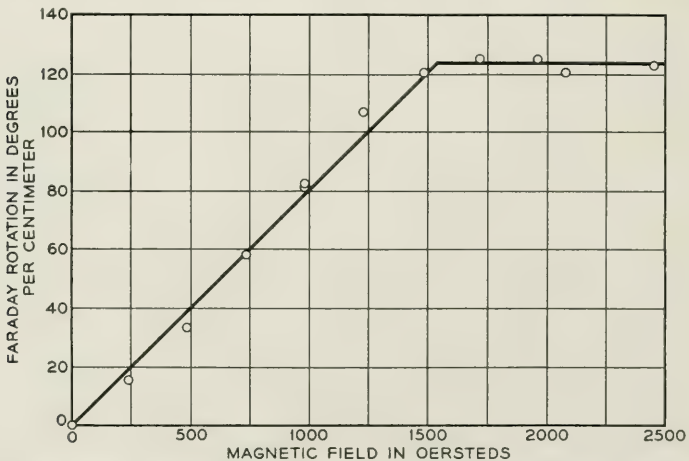


Fig. 7—Angle of rotation versus applied magnetic field for a thin disc of manganese zinc ferrite.

Equation (25). For a particular sample of manganese zinc ferrite, the following measurements were made:

(Sample No. 1)

$$\epsilon' = 17$$

$$\epsilon'' = 24$$

$$4\pi M_{\text{sat}} = 1500 \text{ gauss}$$

Using this data, equation (25) predicts:

$$\frac{\theta}{\ell} = 121.2^\circ/\text{cm}.$$

It is seen in Fig. 6, that the actual measured rotation at saturation is approximately $123^\circ/\text{cm}$. Hence an extremely good agreement with theory has been obtained for this particular sample.

Equation (25) also indicates that the rotation per unit path length should be sensibly independent of frequency within the above approximations. The data are shown in Fig. 8. However, it will be noticed that the frequency difference between these two sets of data is relatively small (3 per cent), and the cumulative experimental error in measuring angles is such that it is difficult to state that the rotation is closer than 1° between the two sets of data. This represents a possible difference of 5 per cent in the rotation for a change of 3 per cent in the frequency. Thus, even though these preliminary data support Equation (25), it cannot be accepted as conclusive evidence until more measurements can be made over a wider band width.

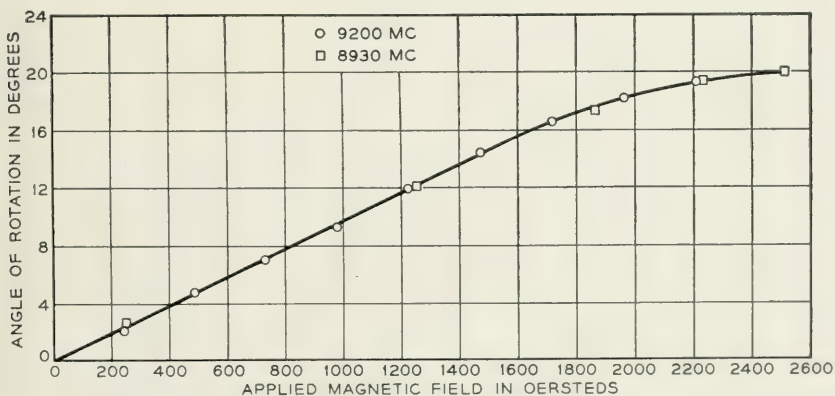


Fig. 8—Dependence of Faraday rotation upon frequency.

The loss characteristics of different ferrites as a function of the applied magnetic field differed distinctly from each other. Some ferrites, such as manganese zinc ferrite showed extremely high loss which was associated with the imaginary part of the dielectric constant. This loss was not affected by the application of a magnetic field but remained substantially constant as the field was applied. However, as the field approached that necessary for ferromagnetic resonance, the total power absorbed by the ferrite increased, since the positive circularly polarized component was almost completely absorbed by the sample. In fact by measuring the ellipticity of the transmitted wave, it is possible to compute the difference between the absorption of the positive and negative circularly polarized components. This has been done for Sample No. 1 and the result is indicated in Fig. 9. If the curve were continued to higher fields, it would represent the shape of the ferromagnetic resonance absorption line.

Some ferrites, such as Ferramic G, showed an almost zero dielectric loss but on the other hand caused an extremely large absorption at 9000 megacycles due to magnetic losses. The major contributions to magnetic loss at this frequency should be either losses associated with a domain wall relaxation or ferromagnetic resonance absorption due to anisotropy fields. Unequivocal data can be obtained by the above techniques to identify which loss is predominant. If the loss were due to domain wall relaxation (or resonance) it would absorb both the negative and positive circularly polarized components equally. Thus as the magnetic field was applied and as the ferrite became saturated, the losses in both components should decrease as the domain walls disappeared. However,

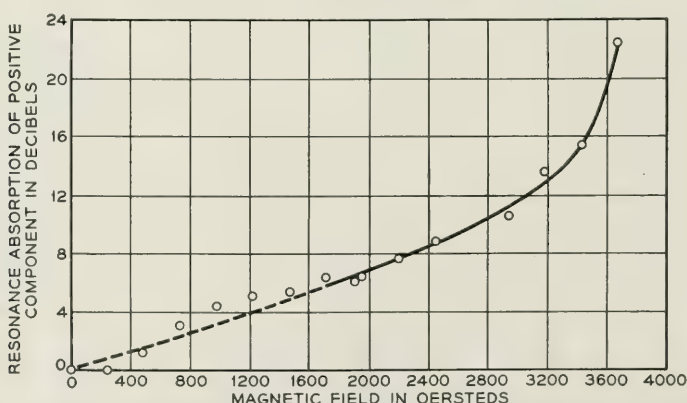


Fig. 9—Ferromagnetic resonance absorption curve determined by measuring the ellipticity of a wave transmitted through a cylinder of ferrite in a waveguide.

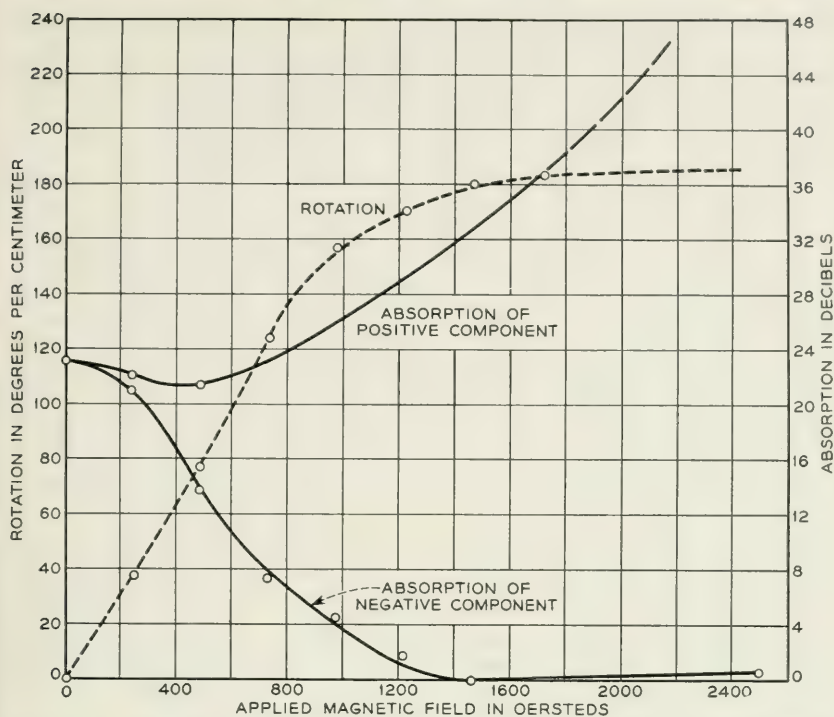


Fig. 10—Rotation of the plane of polarization and absorption of the positive and negative circularly polarized components when an electromagnetic wave (X-band) is propagated through a cylinder of Ferramic G.

if instead the loss is associated with ferromagnetic resonance absorption, the absorption of the positive component should begin to increase as soon as more domains are lined up in a direction where they can absorb the positive component. Thus even before the sample is saturated the absorption of the positive component should be much larger than the absorption of the negative component. Of course, in a polycrystalline sample with a large anisotropy both components can be absorbed by ferromagnetic resonance absorption when the sample is not completely saturated, since the random orientation of the domains which occurred in zero field has not been completely eliminated until complete saturation occurs. Fig. 10 illustrates the rotation per cm path length versus applied magnetic field for a sample of Ferramic G. Superimposed on the same figure are curves showing the absorption of the negative and positive circularly polarized components. It will be noticed that as soon as the sample is saturated, the sample becomes transparent to the negative component but almost completely absorbs the positive component.

Hence the transmitted wave at this point is almost completely circularly polarized, even though the applied magnetic field would indicate that the resonance absorption frequency was far removed from 9000 mc.

Table I gives the data taken on several ferrites at 9000 mc.

APPLICATIONS OF THE FERROMAGNETIC FARADAY EFFECT—THE MICRO-WAVE GYRATOR

As pointed out in the introduction, the Faraday rotation affords an anti-reciprocal phenomenon from which a microwave gyrator can be constructed. Such a gyrator is illustrated in Fig. 11 along with diagrams which help explain its action. Beneath the gyrator are construction lines which indicate the plane of polarization of a wave as it travels through the gyrator in either direction. On each diagram is a dotted sine wave which is for reference purpose only and indicates the constant plane of polarization of an unrotated wave. It is noticed that for propagation from left to right in Fig. 11, the screw rotation introduced by the twisted rectangular guide adds to the 90° rotation given to the wave by the ferrite element making a total rotation of 180° . For a wave travelling in the reverse direction, these two rotations cancel each other, producing a net zero rotation through the complete element. The unique property of the Faraday rotation becomes immediately apparent from this diagram. In the case of the rotation induced by the twisted rectangular guide, the wave rotates in one direction in going from left to right through the twisted section, and rotates in the opposite direction when it transverses the section from right to left. For the case of the rotation induced by the ferrite element, the direction of rotation is indicated by the arrow in the upper figure for either direction of propagation. The important characteristic of the element is the time phase relation between two points such as *A* and *B* in the upper diagram. It is seen with the help of the diagrams illustrating the rotating waves that the field variations are in phase at points *A* and *B* for propagation from left to right, and they are 180° out of phase for propagation from right to left. In other words the transmission line is an integral number of wavelengths long between *A* and *B* for propagation from left to right and is an odd integral number of half wavelengths long for propagation from right to left.

From the above description of the properties of the gyrator, many of its applications in microwave technology become immediately apparent. Before discussing these applications in more detail, however, it is advantageous to introduce standardized terminology and circuit symbols which apply to the gyrator and to other circuit elements derivable from it.

TABLE I

Sample Number	Material	Dimension (cm)	Applied Magnetic Field (oersteds)	Rotation/cm path	Insertion Loss (db)	Ellipticity* (db)	SWR on Input Line (db)
1	BTL $\text{Mn}_{1-\delta}\text{Zn}_{1-\delta}\text{Fe}_2\text{O}_4$	0.447×2.28 (length \times dia.)	0	0	10.0	>50	
			245	15.6	10.3	>50	
			490	33.5	10.0	23.2	
			735	58.2	9.2	15.0	
			980	81.6	9.1	12.1	
			1225	107	9.2	10.9	
			1470	120	10	10.4	
			1715	125	11	9.3	
			1960	123	11.2	9.0	
			2206	121	11.3	7.7	
			2450	123	11.4	6.6	
			2695	—	12.4	5.0	
			2940	—	13.0	3.7	
			3185	—		3.0	
			3675	—		1.4	
2	BTL $\text{Ni}_{1-\delta}\text{Zn}_{1-\delta}\text{Fe}_2\text{O}_4$	1.36×2.28	0	0	0.8	>40	
			245	25	1.9	≈ 40	
			490	44	2.7	≈ 40	
			735	56	2.9	≈ 40	
			980	61	2.7	40	
			1225	68	2.8		
			1715	82	3.3		
			1960	85	4.9		
			2450	118	7.3	0.8	
3	Ferramic A	2.54×0.635	0	0	1.1	>50	0.7
			245	34.9	0.8	"	0.3
			490	43.7	0.8	"	0.3
			735	48.3	0.8	"	0.3
			980	51.1	1.0	"	0.4
			1225	54.0	1.1	"	—
			1715	57.0	1.1	"	—
			1960	60.0	1.9	"	—
			2450	63.0	3.0	35	—
4	Ferramic G	1.77×2.28	2695	64.2	3.7	—	—
			0	0	23.2	$\gg 30$	
			245	38	21.4	23.0	
			490	77	16.7	7.6	
			735	124	12.4	2.1	
			980	157	9.9	1.4	
			1225	170	7.7	0.7	
			1470	180	6.0	0.7	
			3430	c.p.	7.1	0.0	

* Data given is the difference in db between the major and minor components of the elliptically polarized transmitted wave.

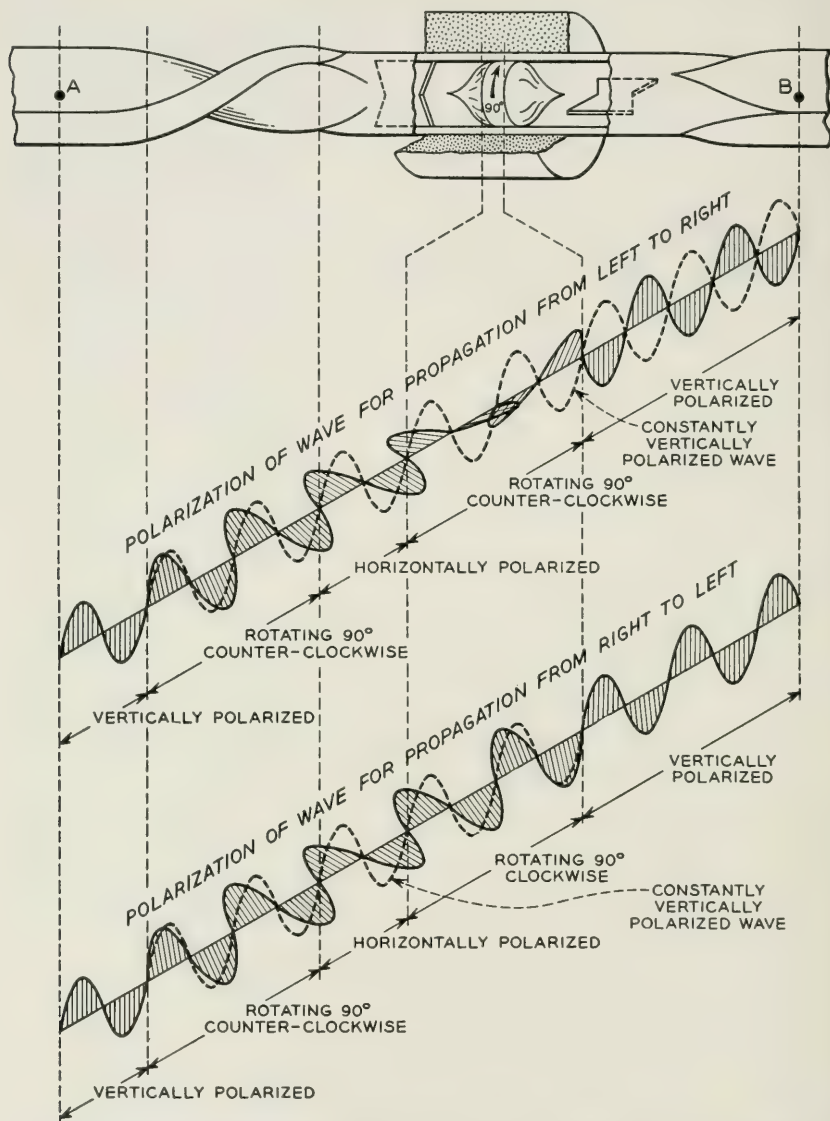


Fig. 11—The microwave gyrator with diagrams which help to explain its operation.

The "active" element of the device, the ferrite cylinder, has been termed a "Faraday Plate."

As was pointed out earlier, the fundamental property of the gyrator is the 180° phase difference introduced between the two directions of propagation through it. Thus the gyrator may be thought of as a four terminal circuit element having no phase shift for one direction of transmission, and having a 180° phase shift for the opposite direction of transmission. A convenient circuit symbol for the gyrator, which indicates this property, is shown in Fig. 12.

If the rectangular waveguides on each side of the Faraday Plate are rotated about their common axis so as to make an angle of 45° with

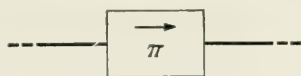


Fig. 12—Circuit symbol for gyrator.

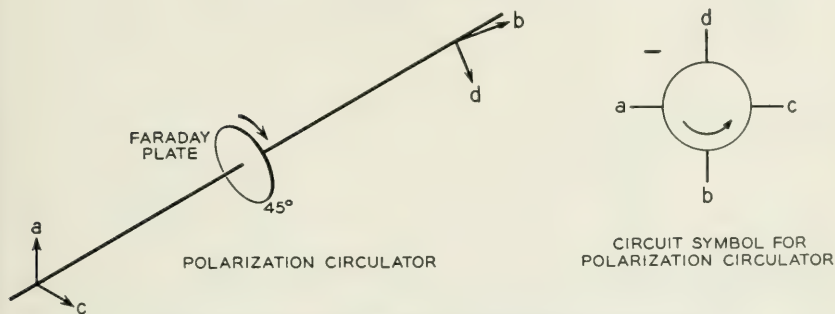


Fig. 13—Schematic diagram of polarization circulator.

each other, then a one-way transmission system can be created which is similar to Lord Rayleigh's one-way transmission system of optics, but with the important difference that this one-way transmission system does not depend upon frequency but is broad band. This one-way transmission system can be used, for example, to isolate the generator or detector from the waveguide in microwave systems. In this application it has the great advantage over the attenuators which are presently used for this purpose in that it can be made practically lossless for the direction of propagation which is desired but the reflected wave will be completely absorbed and hence more complete isolation can be effected.

A more complex and more useful circuit element, than this simple one-way transmission property would at first indicate, is obtained by adding a second connection on each side of the 45° Faraday Plate. It is suggested that this device be called a *polarization circulator*. Thus, the

polarization circulator actually has four output branches corresponding to the two different polarizations at each end. The polarizations of the four output branches are indicated in Fig. 13. It is noticed that power sent into the polarization circulator with polarization a is turned into polarization b , also b is turned into c , c is turned into d , and d is turned into $\text{minus } a$. This property is indicated very clearly by the circuit symbol suggested in Fig. 13, the phase inversion between arms d and a being indicated by the minus sign between the d and a arms.

Another one-way transmission system can be created by combining the gyrator with two-normal hybrids. This combination is indicated in Fig. 14. Since this device has all of the fundamental properties of the

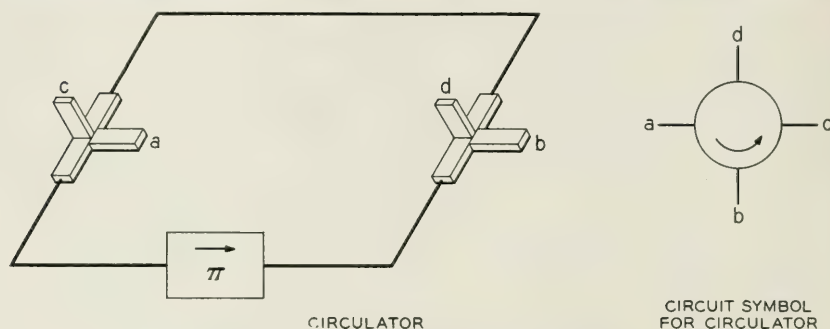


Fig. 14—Schematic diagram of circulator.

polarization circulator with the exception of the phase inversion between arms d and a it is suggested that it be called a "circulator" and the circuit symbol suggested which indicates its properties is also given in Fig. 14.

This list of applications is obviously not complete since it includes only the fundamental elements from which innumerable specific applications can be made.

In addition to the applications discussed above, which depend upon the anti-reciprocal property of the element for their operation there are several simple applications which are based only upon the fact that the amount of rotation can be controlled externally by adjusting the magnetic field. Among these uses are electrically controlled attenuators, modulators, and microwave switches.

ACKNOWLEDGMENTS

The author is indebted to a number of persons for aid in developing this circuit element. In particular, he wishes to thank A. G. Fox for

permission to use his terminology and circuit symbols, and also for the many discussions concerning the properties and uses of these microwave circuit elements. The author is also indebted to S. E. Miller for help in designing the microwave elements and to J. K. Galt for many discussions concerning the theoretical aspects of this paper. The author also wishes to extend his appreciation to J. L. Davis whose able technical assistance made possible the accumulation of much of the data presented in this paper.

APPENDIX

The equation of motion of the magnetization of a ferromagnetic material is:

$$\frac{\partial \vec{M}}{\partial t} = \gamma(\vec{M} \times \vec{H}) - \frac{\gamma\alpha}{|\vec{M}|} [\vec{M} \times (\vec{M} \times \vec{H})] \quad (1)$$

where

\vec{H} = internal magnetic field (oersteds)

$4\pi\vec{M}$ = magnetization of medium (gauss)

α = parameter which measures the magnitude of the damping force on the precessing dipole moment of the sample

γ = gyromagnetic ratio of the electron ($\gamma = ge/2mc$ where g is the Landé g factor for the electron).

If a ferromagnetic material is subjected to a steady magnetic field, H_a , along the z axis and if then an alternating field is applied in an arbitrary direction, Equation (1) must be solved in order to find the behavior of the magnetization of the material. To solve this problem, the following notation is introduced:

$4\pi M_z$ = magnetization of medium in absence of alternating field

H_a = externally applied steady magnetic field (oersteds)

h_x, h_y, h_z = components of applied alternating magnetic field

m_x, m_y, m_z = alternating components of magnetization

h_x^i, h_y^i, H_z^i = components of internal magnetic field

$h_x^i = h_x - N_x m_x$

$h_y^i = h_y - N_y m_y$

$H_z^i = H_a + h_z - N_z(M_z + m_z)$

N_x, N_y, N_z = demagnetizing factors of body.

Hence the magnetic field, H , occurring in Equation (1) is defined by:

$$\vec{H} = h_x \vec{i} + h_y \vec{j} + H_z \vec{k}$$

and

$$\vec{M} = m_x \vec{i} + m_y \vec{j} + (M_z + m_z) \vec{k}$$

In solving Equation (1), an exponential, $\exp [j\omega t]$, time dependence is assumed for the alternating magnetic field and magnetization, and if the following assumption is made:

$$h_x, h_y, h_z \ll H_a$$

it is easily shown that the alternating components of the magnetization of the medium are given by (neglecting terms of the second order in small quantities):

$$\begin{aligned} m_x &= \frac{[\gamma^2 M_z H_z^i (1 + \alpha^2) + j\gamma \alpha M_z \omega] h_x^i - j\gamma M_z \omega h_y^i}{\gamma^2 H_z^{i2} (1 + \alpha^2) - \omega^2 + j[2\omega \gamma \alpha H_z^i]} \\ m_y &= \frac{[\gamma^2 M_z H_z^i (1 + \alpha^2) + j\gamma \alpha M_z \omega] h_y^i + j\gamma M_z \omega h_x^i}{\gamma^2 H_z^{i2} (1 + \alpha^2) - \omega^2 + j[2\omega \gamma \alpha H_z^i]} \\ m_z &= 0 \end{aligned} \quad (2)$$

where:

$$j = \sqrt{-1}$$

Since

$$\vec{b} = \vec{h}^i + 4\pi \vec{m}, \quad (3)$$

it is possible by means of Equations (2) and (3) to find the relation between the alternating flux density \vec{b} and the internal alternating field \vec{h}^i . If the ferromagnetic body is considered as being infinite, the internal fields and applied fields are equal. Hence, for this case:

$$\begin{aligned} b_x &= \mu h_x - jK h_y \\ b_y &= jK h_x + \mu h_y \\ b_z &= h_z \end{aligned} \quad (4)$$

where:

$$\begin{aligned} \mu &= \mu' - j\mu'' \\ K &= K' - jK'' \end{aligned}$$

and:

$$\mu' = 1 + \frac{[\gamma^2 H_a^2 (1 + \alpha^2) - \omega^2][4\pi M_z \gamma^2 H_a (1 + \alpha^2)] + 8\pi M_z \omega^2 \gamma^2 \alpha^2 H_a}{[\gamma^2 H_a^2 (1 + \alpha^2) - \omega^2]^2 + 4\omega^2 \gamma^2 \alpha^2 H_a^2}$$

$$K' = \frac{4\pi M_z \gamma \omega [\gamma^2 H_a^2 (1 + \alpha^2) - \omega^2]}{[\gamma^2 H_a^2 (1 + \alpha^2) - \omega^2]^2 + 4\omega^2 \gamma^2 \alpha^2 H_a^2}$$

$$K'' = \frac{8\pi M_z \omega^2 \gamma^2 \alpha H_a}{[\gamma^2 H_a^2 (1 + \alpha^2) - \omega^2]^2 + 4\omega^2 \gamma^2 \alpha^2 H_a^2}$$

$$\mu'' = \frac{4\pi M_z \gamma \alpha \omega [\gamma^2 H_a^2 (1 + \alpha^2) + \omega^2]}{[\gamma^2 H_a^2 (1 + \alpha^2) - \omega^2]^2 + 4\omega^2 \gamma^2 \alpha^2 H_a^2}$$

In order to find the behavior of a wave being propagated in this medium, it is necessary to find a solution to Maxwell's equations which are consistent with the above set of equations and in which, b , h , E , and D are of the following form:

$$\begin{aligned}\vec{b} &= \vec{b}_0 \exp [j\omega t - \Gamma(\vec{n} \cdot \vec{r})] \\ \vec{h} &= \vec{h}_0 \exp [j\omega t - \Gamma(\vec{n} \cdot \vec{r})] \\ \vec{E} &= \vec{E}_0 \exp [j\omega t - \Gamma(\vec{n} \cdot \vec{r})] \\ \vec{D} &= \vec{D}_0 \exp [j\omega t - \Gamma(\vec{n} \cdot \vec{r})]\end{aligned}\tag{5}$$

where \vec{E}_0 , and \vec{h}_0 are complex vector functions of the coordinates and which satisfy the boundary conditions imposed by the waveguide. Further:

\vec{n} = unit vector in the direction of propagation

Γ = propagation constant

Maxwell's equations are:

$$\begin{aligned}\nabla \times \vec{E} &= -\frac{1}{c} \frac{\partial \vec{b}}{\partial t} \\ \nabla \times \vec{h} &= \frac{1}{c} \frac{\partial \vec{D}}{\partial t}\end{aligned}\tag{6}$$

Inserting the values given in Equations (5), these become:

$$\nabla \times \vec{E}_0 - \Gamma(\vec{n} \times \vec{E}_0) = \frac{-j\omega \vec{b}_0}{c}\tag{7}$$

$$\nabla \times \vec{h}_0 - \Gamma(\vec{n} \times \vec{h}_0) = \frac{j\omega \epsilon \vec{E}_0}{c}\tag{8}$$

which can be combined to:

$$\frac{\omega^2 \epsilon}{c^2} \vec{b}_0 = \nabla \times (\nabla \times \vec{h}_0 - \Gamma \vec{n} \times \vec{h}_0) - \Gamma \vec{n} \times (\nabla \times \vec{h}_0 - \Gamma \vec{n} \times \vec{h}_0) \quad (9)$$

Writing Equation (9) in component form gives:

$$\begin{aligned} \frac{\omega^2 \epsilon}{c^2} b_x &= \frac{\partial^2 h_y}{\partial y \partial x} + \frac{\partial^2 h_z}{\partial z \partial x} - \frac{\partial^2 h_x}{\partial y^2} - \frac{\partial^2 h_x}{\partial z^2} - \Gamma n_x \left(\frac{\partial h_y}{\partial y} + \frac{\partial h_z}{\partial z} \right) \\ &+ 2\Gamma \left(n_y \frac{\partial h_x}{\partial y} + n_z \frac{\partial h_x}{\partial z} \right) - \Gamma \left(n_y \frac{\partial h_y}{\partial x} + n_z \frac{\partial h_z}{\partial x} \right) \\ &+ \Gamma^2 n_x (n_x h_x + n_y h_y + n_z h_z) - \Gamma^2 h_x \end{aligned} \quad (10)$$

$$\begin{aligned} \frac{\omega^2 \epsilon}{c^2} b_y &= \frac{\partial^2 h_x}{\partial y \partial x} + \frac{\partial^2 h_z}{\partial y \partial z} - \frac{\partial^2 h_y}{\partial x^2} - \frac{\partial^2 h_y}{\partial z^2} - \Gamma n_y \left(\frac{\partial h_x}{\partial x} + \frac{\partial h_z}{\partial z} \right) \\ &+ 2\Gamma \left(n_x \frac{\partial h_y}{\partial x} + n_z \frac{\partial h_y}{\partial z} \right) - \Gamma \left(n_x \frac{\partial h_x}{\partial y} + n_z \frac{\partial h_z}{\partial y} \right) \\ &+ \Gamma^2 n_y (n_x h_x + n_y h_y + n_z h_z) - \Gamma^2 h_y \end{aligned} \quad (11)$$

$$\begin{aligned} \frac{\omega^2 \epsilon}{c^2} b_z &= \frac{\partial^2 h_x}{\partial z \partial x} + \frac{\partial^2 h_y}{\partial z \partial y} - \frac{\partial^2 h_z}{\partial x^2} - \frac{\partial^2 h_z}{\partial y^2} \\ &- \Gamma n_z \left(\frac{\partial h_x}{\partial x} + \frac{\partial h_y}{\partial y} \right) + 2\Gamma \left(n_x \frac{\partial h_z}{\partial x} + n_y \frac{\partial h_z}{\partial y} \right) \\ &- \Gamma \left(n_x \frac{\partial h_x}{\partial z} + n_y \frac{\partial h_y}{\partial z} \right) + \Gamma^2 n_z (n_x h_x + n_y h_y + n_z h_z) \\ &- \Gamma^2 h_z \end{aligned} \quad (12)$$

where the subscript 0 has been dropped from all components for convenience.

If the wave in question is an infinite plane wave being propagated along the z axis, then:

$$n_x = n_y = 0, \quad n_z = 1$$

and the components of \vec{h}_0 are constants, and h_z is zero. For this particular case, Equations (10), (11) and (12) become:

$$\frac{\omega^2 \epsilon}{c^2} b_x = -\Gamma^2 h_x \quad (13)$$

$$\frac{\omega^2 \epsilon}{c^2} b_y = -\Gamma^2 h_y \quad (14)$$

Equations (13) and (14) are general differential equations derivable from Maxwell's equations and do not yet contain the properties of any particular medium. In order to find the behavior of a wave travelling through an infinite ferromagnetic medium which is magnetized along the direction of propagation, it is necessary to combine these equations with Equations (4) which describe the relation between b and h in the medium.

This gives:

$$(\mu h_x - jKh_y) \frac{\omega^2 \epsilon}{c^2} = -\Gamma^2 h_x \quad (15)$$

$$(\mu h_y + jKh_x) \frac{\omega^2 \epsilon}{c^2} = -\Gamma^2 h_y \quad (16)$$

The only possible solution to this set of equations is a circularly polarized wave where:

$$h_x = \pm jh_y$$

The positive sign above represents a so-called positive circularly polarized wave and the negative sign a negative circularly polarized wave. The propagation constants for these waves is given by

$$\Gamma_{\pm} = \frac{j\omega}{c} \sqrt{\epsilon(\mu \pm K)} \quad (17)$$

REFERENCES

1. B. D. H. Tellegen, *Philips Research Reports*, **3**, pp. 81-101 (1948); **3**, pp. 321-337 (1948); **4**, pp. 31-37 (1949); **4**, pp. 366-369 (1949).
2. A. G. Fox, unpublished memoranda, Bell Telephone Laboratories.
3. E. M. McMillan, *J. Acous. Soc. of Am.*, **18**, pp. 344-347 (1946).
4. J. W. Miles, *J. Acous. Soc. of Am.*, **19**, pp. 910-913 (1947).
5. F. Bloch, *Phys. Rev.*, **70**, pp. 460-485 (1946).
6. H. G. Beljers and J. L. Snoek, "Gyromagnetic Phenomena Occurring Within Ferrites", *Philips Tech. Rev.*, **11**, May, 1950, pp. 313-322.
7. E. M. McMillan, *J. Acous. Soc. of Am.*, **19**, p. 922 (1947) and H. B. G. Casimir, *Nuovo Cimento*, Vol. VI, Series IX (1949).
8. W. P. Mason, unpublished memoranda, Bell Telephone Laboratories.
9. Rayleigh, *Nature*, Vol. 64.
10. H. König, *Optik*, **3**, pp. 101-119 (1948).
11. D. Polder, *Phil. Mag.*, **40**, pp. 99-115 (1949).
12. W. A. Yager, J. K. Galt, et al. *Phys. Rev.*, **80**, pp. 744-748 (1950).

Dialing Habits of Telephone Customers

BY CHARLES CLOS AND ROGER I. WILKINSON

(Manuscript received October 3, 1951)

This paper considers the behavior of customers waiting to dial calls, when dial tone is delayed. Tests were made in a panel dial central office, from which were determined: relationship between load carried by a group of line finders and the resultant dial tone delay; measures, by classes of service, of the magnitude of the generalized trunking formula's "j" factor describing the degree to which customers wait when dial tone is delayed; comparisons of observed and theoretical distributions of the number of simultaneous calls on line finder groups; and statistical accounts of the actions of customers when dial tone is delayed.

Following World War II the conversion of great quantities of manual telephone equipment to dial, and the addition of large numbers of new telephones, mostly dial, in the Bell System has directed increasing attention to those service problems peculiar to automatic operation. These problems concern chiefly the provision of adequate amounts of equipment to give satisfactory service at all times. One of the important factors affecting the amount of equipment needed is the action of the customers themselves when their calls are momentarily blocked due to these equipment shortages. The actions of subscribers whose calls are blocked due to a shortage of trunk equipment have been reported previously.¹ This paper considers the behavior of subscribers, waiting to dial calls, when dial tone is delayed.

During 1949, Bell Telephone Laboratories conducted a series of tests at the New York Telephone Company's Sterling-3 panel dial central office in Brooklyn, N. Y., with the object of increasing the knowledge available regarding subscribers' actions and their effects when dial tone is delayed.

The following principal results were obtained from the Sterling-3 tests:

1. The relationship between the load carried by a group of line finders and the resultant dial tone delay.
2. Measures, by classes of service, of the magnitude of the general-

¹ Charles Clos, "An Aspect of the Dialing Behavior of Subscribers and Its Effect on the Trunk Plant," *Bell System Tech. J.*, **27**, July 1948.

ized trunking formula's "j" factor describing the degree to which customers wait when dial tone is not immediate.²

3. Comparisons of observed and theoretical distributions of the numbers of simultaneous calls on the line finder groups.

4. Statistical accounts of the actions of subscribers when dial tone is delayed.

GENERAL PLAN OF THE TESTS

The general plan of the tests was developed around two specially constructed devices:

1. A 100-pen recorder capable of recording observations continuously for two hours and sensitive enough to record individual dial pulses.

2. A speed of dial tone measuring set comprising a means for manually originating test calls, and an electric timer which automatically stopped when dial tone was received.

Tests were conducted on a weekly schedule, one or two line finder groups being studied each week. Two 100-pen recorder tapes were run daily. The first was run from 10 A.M. to noon for message rate individual, flat rate individual and message rate two-party classes of service to include the morning busy periods for these customers. For coin customers the first tape was run for two hours during the noon coin busy period. The second tape was run in the afternoon from 2 to 4 P.M. for all classes of service except coin. For the coin class the second tape was run for two hours during the early evening coin busy period. Three message rate two-party tapes were run in the early evening busy period, and two coin tapes were run in the afternoons when World Series baseball games were being played in Brooklyn. A summary of the number of line finder groups observed, the number of tapes taken, the number of line finders made available and the maximum per cent dial tone delays over three seconds are given in Table I.

Except for the morning runs on the message rate individual line groups where an effort was made to maintain a fixed number of twenty line finders throughout, the number of line finders made available was selected by close observation of the flow of traffic. All studies were by half hours during which the number of line finders was held constant as far as possible. At the end of a half-hour period the number of line

² This formula for both finite and infinite sources was developed by R. I. Wilkinson in 1930, and appeared in the 1936 Bell Telephone Laboratories Out-of-Hour Course "The Theory of Probability as Applied to Telephone Trunking Problems." This formula for infinite sources was also developed by Conny Palm and appeared in "Etude des delais d'attente" in Erickson Technics—No. 2—1937.

finders left in service was adjusted in order to obtain a reasonable number of dial tone delays in the next half hour without producing a severe reaction from the subscribers served by the line groups under study.

The data recorded on the tapes showed continuously the busy or idle conditions of certain circuits associated with the line groups under study. In some cases the receipt of dial pulses and the operation of registers were also recorded. These circuits included line finders, a few subscriber lines, trip circuit sub-groups, the all trunks busy register, the peg count register and the speed of dial tone measuring device. In

TABLE I

	Number of Line Finder Groups Observed	Number of Tapes	Number of Line Finders Made Available	Maximum Percent Dial Tone Delays Over 3 Seconds
Message rate individual.....	11			
Morning tapes.....		25	19 to 20	53.3
Afternoon tapes.....		29	10 to 15	55.6
Message rate two-party.....	3			
Morning tapes.....		8	18 to 19	40.0
Afternoon tapes.....		12	10 to 12	88.6
Evening tapes.....		3	19 to 21	71.2
Flat rate individual.....	2			
Morning tapes.....		9	14 to 33	48.9
Afternoon tapes.....		10	10 to 17	35.6
Coin.....	3			
Morning tapes.....		18	30 to 39	71.1
Afternoon tapes.....		2	30 to 39	35.6
Evening tapes.....		10	25 to 39	68.9

addition the busy and idle conditions of a sample of senders was observed in order to note the general load level on the senders.

RELATIONSHIP BETWEEN LOAD CARRIED AND PER CENT DIAL TONE DELAY

One of the principal objectives of these tests was to establish as far as possible the relationship between the average load carried by a line-finder group and the corresponding dial tone service when there is a shortage of line finders but not of senders.³ The average load carried was obtained by making a switch count every thirty seconds of the number of line finders busy as indicated by the 100-pen recorder tape. The dial tone tests were made with the speed of dial tone measuring device. Forty-five dial tone tests were made each half hour for each

³ The restriction of avoiding dial-tone delays due to a sender shortage was to eliminate a factor external to the line-finder group and to make the results of the tests applicable to both common-control and non-common-control type dial systems.

line group studied. Additional dial tone tests were made on all other line groups in the office as a check that the delays experienced on the line groups under study were not due to a sender shortage. The sender data on the tapes were also used for this purpose. Figs. 1(a) to 1(d), and 2(a) to 2(d), inclusive, show for various amounts of load carried, the per cent of dial tone tests encountering delays greater than three and greater than ten seconds for half-hour study periods for the most frequent number of line finders in the tests for each class of service.

Plotted on each of these figures is a theoretical fitting dial tone tester delay curve computed for the indicated dial tone delays and for the following j factor values in the generalized trunking formula, determined in a manner to be explained later:

Class of Service	j factor
MRI	6.6
MR 2-party	5.8
FRI	6.5
Coin	2.1

To indicate the effect of varying j , several curves have been added to Fig. 1(a). Selections of curves for $j = 0, 1$ and ∞ (which correspond to the three commonly used infinite source congestion formulae, Erlang C, Poisson and Erlang B when adapted to the tester's delay problem) are shown. It is clear that with the wide differences in delays which they give for specified loads carried, it is highly desirable to select that j formula for engineering use which most nearly describes the customer actions in any situation being dealt with. In the field of curves shown on Fig. 1(a), the one labelled $j = 6.6$ was derived in a logical manner from the data, and shows an agreeably satisfactory fit. For example, during a heavy load period when, say, 20 per cent of a dial tone tester's calls are meeting delays greater than 3 seconds, an actual average load of 16.6 erlangs (as shown by the $j = 6.6$ curve) would likely be carried. (Load in erlangs equals average number of simultaneous calls.) The Erlang C ($j = 0$) and Poisson ($j = 1$) theories would indicate the presence of loads of 15.6 and 16.0 erlangs, respectively, figures clearly too small for the circumstances shown by the data of Fig. 1(a). On the other hand, use of the Erlang B ($j = \infty$) theory would predict a considerably larger load carried, about 17.9 erlangs, than one would probably be justified in assuming here for engineering purposes.

By grouping the dial tone delay data by bands of load carried, relationships of per cent of test calls encountering varying dial tone delays

up to twelve seconds were obtained. These data are shown on Figs. 3 to 6. Fig. 3 is for the message rate individual class of service with twenty line finders available. The data on this figure correspond to those on Figs. 1(a) and 1(b). Fig. 4 is for the message rate two-party class of service with ten line finders and corresponds to Figs. 1(c) and 1(d). Fig. 5 is for the flat rate class of service with ten line finders and cor-

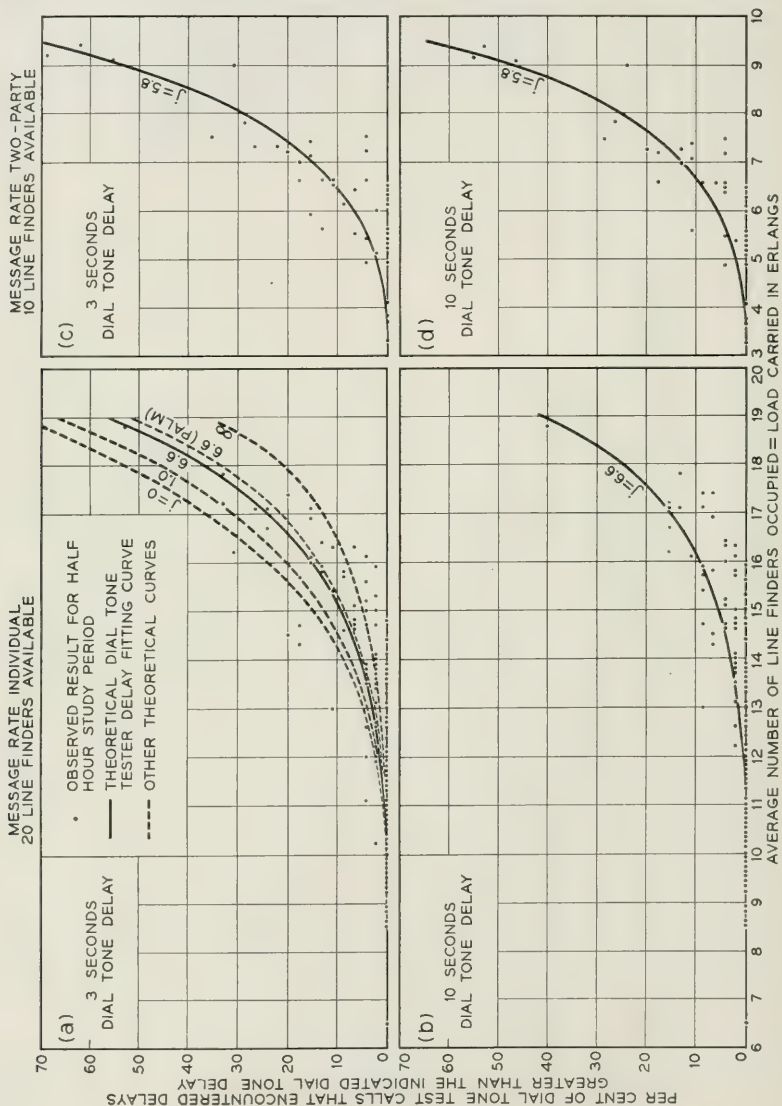


Fig. 1.—Results of dial tone tests.

responds to Figs. 2(a) and 2(b). Fig. 6 is for the coin class of service with 34 line finders and corresponds to Figs. 2(c) and 2(d).

Plotted on Figs. 3 to 6 are theoretical fitting dial tone tester delay curves, curves A, determined by means of the following formulae:

1. The generalized trunking formula for determining the proportion of calls that encounter congestion, i.e., find all line finders busy.

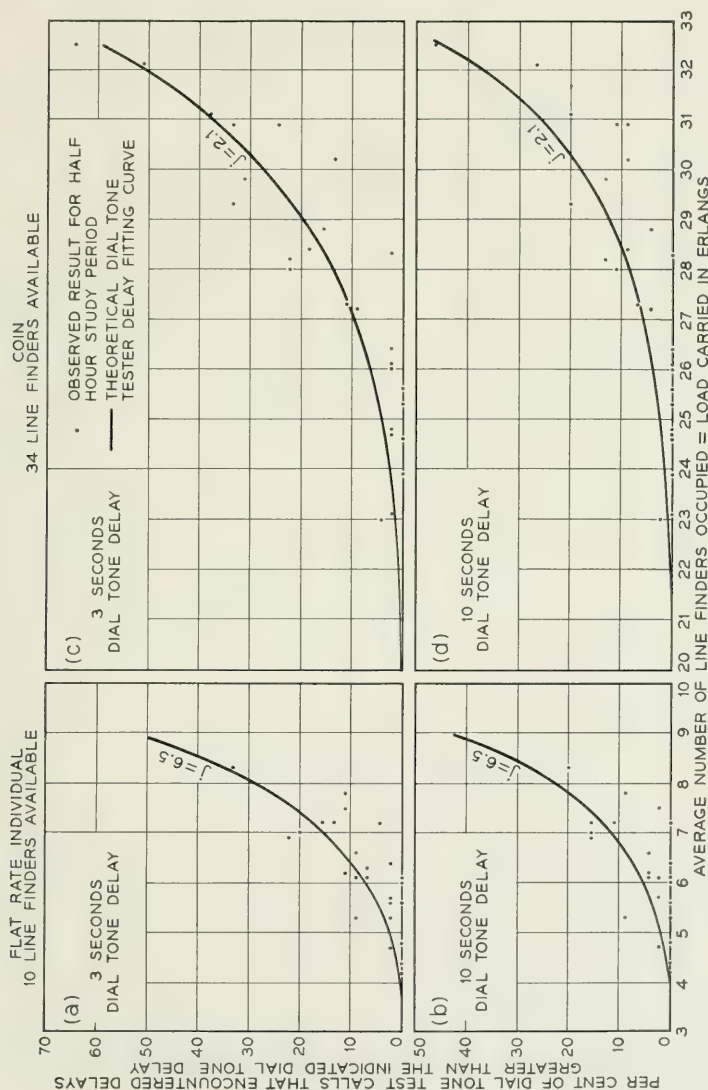


Fig. 2—Results of dial tone tests.

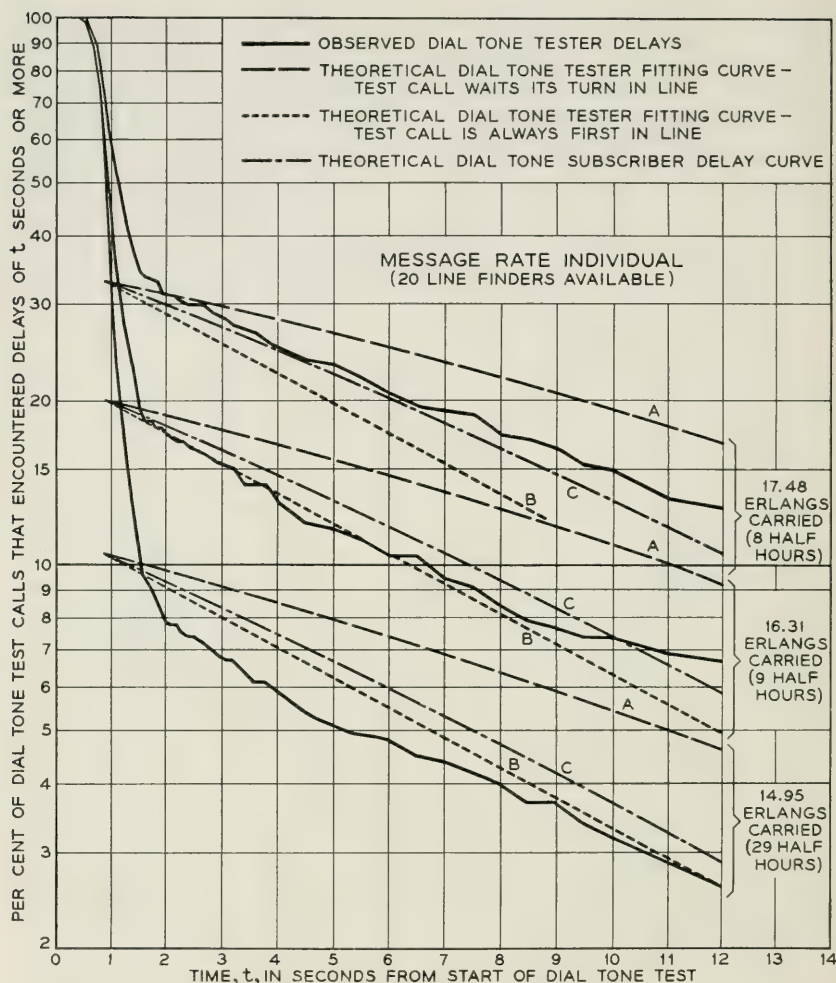


Fig. 3—Results of dial tone tests.

2. A delay formula⁴ for determining the proportion of those test calls encountering congestion which will have a delay in obtaining dial tone of at least time t .

Some of the theoretical aspects of these formulae are considered in the two sections that follow.

GENERALIZED TRUNKING FORMULA

The generalized trunking formula combines in one expression the various assumptions underlying the Erlang B, Poisson, and Erlang C

⁴ A development by John Riordan paralleling that of Conny Palm's.

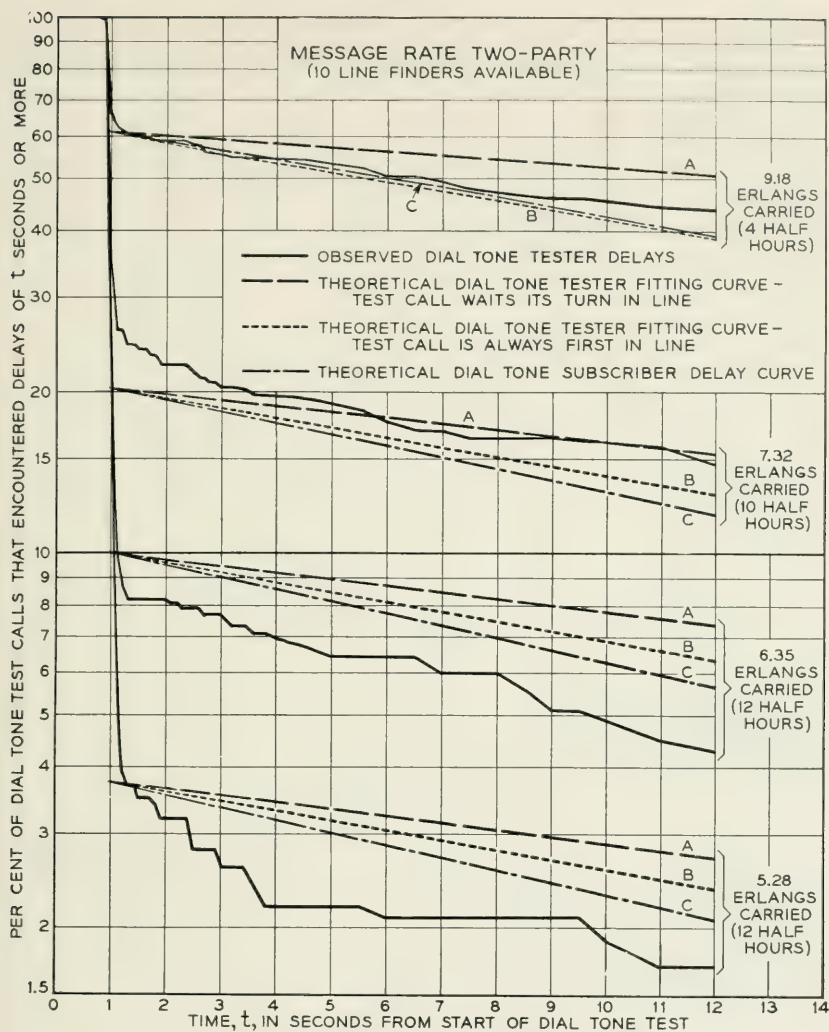


Fig. 4—Results of dial tone tests.

trunking formulae and a large field of intermediate assumptions regarding the disposition of calls which fail to obtain immediate service. These assumptions are given in Table II on page 42.

One method for developing the generalized trunking formula⁵ is to consider the probabilities of the existence of certain states and to determine the number of transitions during some convenient interval of

⁵ In this article only the unlimited sources trunking formula is considered.

time from one state to another. By equating certain of these transitions, a series of simultaneous equations evolves, which when solved yields one overall expression. Of interest is the development of the transition equations. Thus for a case of c line finders arranged in a simple group, let $f(x)$ represent the probability that x (where $x < c$) line finders are occupied and $f(x + 1)$ represent the probability that $x + 1$ line finders are occupied. Let n be the average number of calls offered to the line finders during a long interval of time, T . Let T be the unit of time and h be the average holding time per call measured in terms of T . Over a

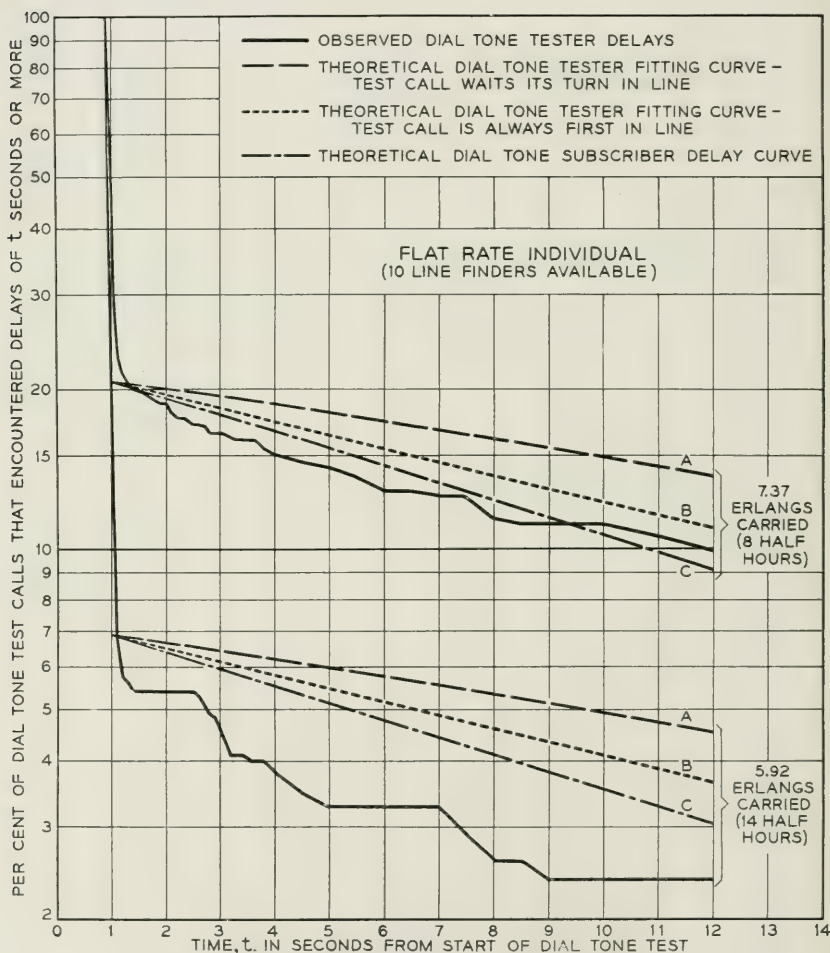


Fig. 5—Results of dial tone tests.

period T the number of transits from state x to state $x + 1$ must equal (or differ by no more than one) the number of transits in the reverse direction. That is:

$$nf(x) = \frac{x+1}{h} f(x+1) \quad (1)$$

It may be noted that $nh = a$, where a is the average offered traffic load in erlangs.

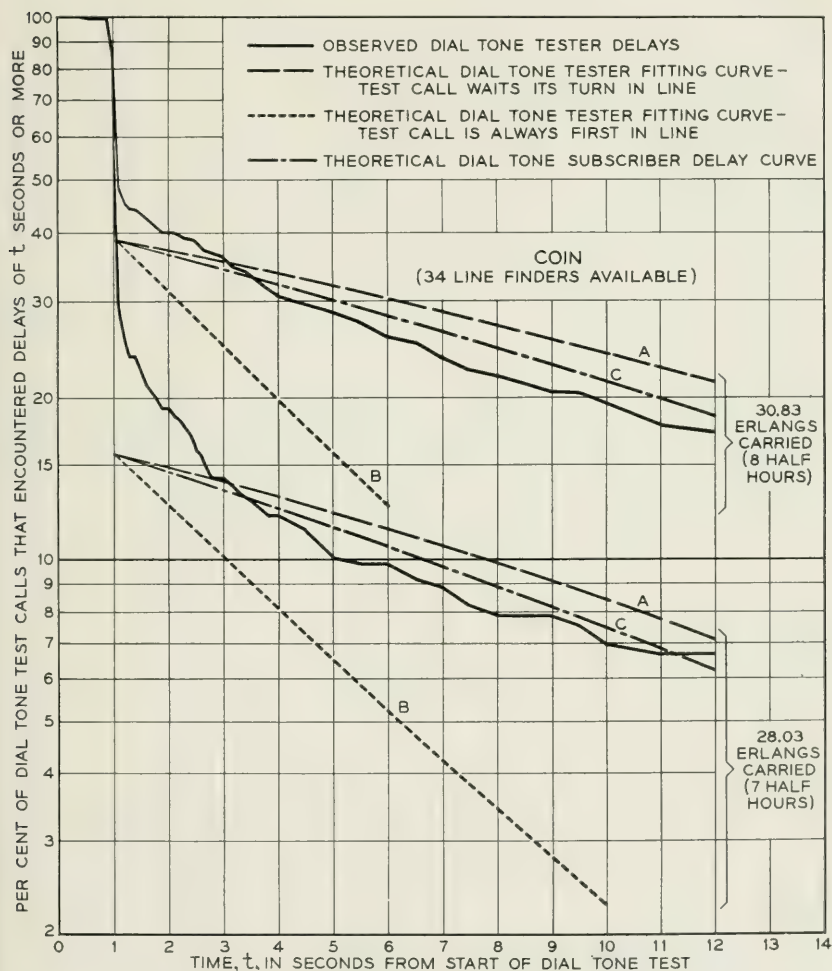


Fig. 6—Results of dial tone tests.

TABLE II

Formula	Assumption Concerning the Disposal of Calls that do not Obtain a Line Finder Immediately
Generalized.....	Waiting calls are cleared out at a rate j times the rate at which calls are terminated when served by the line finders.
Erlang B.....	Calls are cleared out of the system immediately, that is no calls wait ($j = \infty$).
Poisson.....	Waiting calls are cleared out at a rate equal to that with which calls are terminated when served by the line finders ($j = 1$).
Erlang C.....	Calls wait until served ($j = 0$).

When $x \geq c$ a new situation is encountered. " c " calls are engaged in conversation and $x - c$ calls are waiting for service. If the waiting calls are forced to wait for an unduly long period of time so that in effect they are being denied service, it can be expected that they will wait for some average period, say H , and then abandon their attempts. On this basis the corresponding equation is:

$$nf(x) = \frac{c}{h} f(x+1) + \frac{(x+1-c)}{H} f(x+1) \quad (2)$$

It has been assumed in the above equations that the distribution of the holding times is exponential, an assumption which is found in most local systems to be reasonably justified. The distribution of the waiting times is also taken to be exponential. By introducing a factor j , where $j = h/H$, equation (2) can be written in the simpler form:

$$af(x) = [c + j(x+1-c)]f(x+1) \quad (3)$$

Solving this system of simultaneous equations, we obtain: when $x < c$,

$$f(x) = \frac{a^x}{x!} f(0) \quad (4)$$

when $x \geq c$,

$$f(x) = \frac{a^x}{c!(c+j)(c+2j) \cdots [c+(x-c)j]} f(0) \quad (5)$$

where

$$f(0) = \left(\sum_{x=0}^{x=c} \frac{a^x}{x!} + \sum_{x=c+1}^{x=\infty} \frac{a^x}{c!(c+j)(c+2j) \cdots [c+(x-c)j]} \right)^{-1} \quad (6)$$

The probability of a call encountering congestion, which is equivalent to the probability of a call having a delay greater than zero units of time is:

$$P(>0) = \sum_{x=c}^{x=\infty} f(x) \quad (7)$$

DELAY FORMULA FOR THE DIAL TONE TESTER

The probability that a dial tone test call encounters congestion is given by expression (7). Once a test call has encountered congestion it will experience a delay depending upon a number of variables. The assumptions underlying the dial tone tester formula are:

1. A dial tone test call when encountering a delay waits until served.
2. A dial tone test call does not add to the load offered and carried by the line finders.
3. Upon encountering a delay, a dial tone test call is served in the order of its arrival with respect to all other waiting calls. For example, if the test call finds three other calls waiting, it waits fourth in line.

Under the third assumption as calls drop out, due to conversations terminating on the occupied line finders or due to waiting calls abandoning their requests for service, the test call advances from an initial position of say fourth in line to third in line, then to second, then to first in line, and finally is served. The overall delay distribution of the test calls depends therefore upon the number of calls they find waiting ahead of them. The delay distribution for each such number must be weighted by the probability of its occurrence in order to obtain the overall distribution. The delay distribution for a test call which finds zero calls waiting is:

$$p_0(>t) = \exp(-ct/jH) \quad (8)$$

The probability is $f(c)$ that a call made at random will find all line finders busy with no calls waiting. Hence the weighted delay distribution $P_0(>t)$, is:

$$P_0(>t) = f(c)p_0(>t) = f(c) \exp(-ct/jH) \quad (9)$$

The delay distribution for a test call which finds one call waiting ahead of it is:

$$p_1(>t) = [1 + c/j - (c/j) \exp(-t/H)] \exp(-ct/jH) \quad (10)$$

The probability is $f(c+1)$ that a call made at random will find all

line finders busy and one call waiting. Hence the weighted delay distribution, $P_1(>t)$, is:

$$P_1(>t) = f(c+1)[1 + c/j - (c/j) \exp(-t/H)] \exp(-ct/jH) \quad (11)$$

In the general case $p_n(>t)$ is given by the following formula:

$$p_n(>t) = F_{n+1}(t) \exp(t/H) \quad (12)$$

where $F_{n+1}(t)$ is given by Conny Palm.⁶ The over-all delay distribution is then:

$$P(>t) = P_0(>t) + P_1(>t) + P_2(>t) + \dots \quad (13)$$

By making appropriate substitutions and summing the result, expression (13) becomes:

$$P(>t) = \frac{a^c}{c!} f(0) \left[1 + \frac{a \exp(-t/H)}{c+j} + \frac{a^2 \exp(-2t/H)}{(c+j)(c+2j)} + \dots \right] \exp \left[-ct/jH + (a/j)[1 - \exp(-t/H)] \right] \quad (14)$$

Expression (14) is equivalent to that of Riordan involving two incomplete gamma functions as follows:

$$P(>t) = P(>0) \frac{\gamma[c/j, (a/j) \exp(-t/H)]}{\gamma(c/j, a/j)} \quad (15)$$

where the incomplete gamma function,

$$\gamma(N, x) = \int_0^x x^{N-1} e^{-x} dx \quad (16)$$

The theoretical dial tone tester delay curves shown on Figs. 1(a) to 1(d), 2(a) to 2(d), and 3 to 6 were computed from expression (14), using the following values of j and H for the classes of service studied, these values being determined in a manner explained later:

Class of Service	j factor	H
MRI	6.6	24 seconds
MR 2-party	5.8	42 seconds
FRI	6.5	27 seconds
Coin	2.1	74 seconds

On Figs. 1(a), 1(c), 2(a), and 2(e), which show the per cent of dial tone tests encountering delays greater than three seconds for various amounts

⁶ Equation 53, loc. cit.

of load carried, it may be noted that most of the theoretical dial tone tester delay curves are in close agreement with the observed data, with a tendency perhaps to be slightly high. On Figs. 1(b), 1(d), 2(b), and 2(d), which are for dial tone delays greater than ten seconds, it may be noted that the theoretical curves have a slightly stronger tendency to lie on the high side of the observed data. On Figs. 3 to 6 the theoretical dial tone tester delay, curves A, again lie in the proximity of the curves of the observed data, with a tendency to lie higher than these latter curves, especially at the ends where the dial tone delays are greatest. Among the factors which account for this discrepancy are:

1. A feature is present in panel line finder circuits for momentarily releasing trip circuits with waiting calls to prevent the orphaning of calls under certain trouble conditions. The release occurs after a call has been waiting from 5 to 12 seconds and reoccurs every 7 seconds thereafter. When such a release occurs the call yields whatever waiting preference it may have had to a subsequently placed call which is not yet affected by such a release. The dial tone test calls did not wait beyond 12 seconds. Hence for these test calls there was only one possibility of such a release and for many of them the release occurred near the end of their waiting period. Hence they were more likely to gain preference over other calls than to lose their preference.

2. Subscribers while waiting for dial tone frequently become impatient and proceed to flash (move their switchhook up and down). While flashing, a subscriber may lose preference to a subsequently placed test call (the latter of course does not flash).

3. Many subscribers fail to observe dial tone and proceed to dial. During such dialing, a subscriber may lose preference to a subsequently placed test call.

4. Line finders serve a large proportion of call attempts of short holding time whose presence may militate against the occurrence of the longer delays. In connection with the measurement of the j factor, the following proportions of call attempts and average holding times were noted on which no dialing occurred or where no more than two digits were dialed.

Class of Service	Proportion of Attempts with No Peg Counts	Average Holding Time
MRI	35.2%*	4.4 seconds*
MR 2-Pty.	33.3%*	5.8 seconds*
FRI	25.1%	5.8 seconds
Coin	9.8%	8.3 seconds

* Partly estimated

The individual contributions of these four factors to discrepancies between theory and observation are not easy to assess. The first three explain a tendency for test calls to get ahead of calls already waiting for dial tone. On Figs. 3 to 6, inclusive, additional theoretical dial tone delay curves, curves B, for the case where a dial tone tester always gets first in line are shown. Even these curves tend to lie above the curves of the observed data on Figs. 4 and 5 where ten line finders were available; they more nearly agree with the observed data on Fig. 3 where twenty line finders were available, and they lie below the observed data on Fig. 6 where 34 line finders were available. This is an indication of the fact that with higher traffic loads (which occurred on the larger line finder groups) a test call will encounter more competition from other calls and therefore will have a lesser chance of gaining precedence over all of the other calls. The fourth factor indicates that the call attempts served on line finders consist of two distinct holding time universes and not just one, as was assumed in the development of the dial tone tester formula. The effect of the presence of both a short and long holding time universe of calls would be to introduce a change of slope in the delay curves which may be seen in Figs. 3 to 6 to be at about $t = 4$ seconds. There is reason to believe that the same cause may have been responsible for the tendency of the observed delay curves to fall away from the theoretical at the lower levels of load carried.

Due to the reasons given above and to the fact that the dial tone delay observations were made by the test call method, the above results may not directly describe service from the customer's point of view. Conny Palm has developed the following formula which gives a slightly different measure of customers' dial tone service. It indicates the proportion of calls which have neither received dial tone nor have dropped out at time t .

$$P(>t) = P(>0) \frac{\gamma[c/j, (a/j) \exp(-t/H)]}{\gamma(c/j, a/j)} \exp(-t/H) \quad (17)$$

Curves for this formula are shown plotted on Fig. 1(a) and at C on Figs. 3 to 6. They are quite close in many cases to the observed dial tone tester results. It would appear that a sufficiently good estimate of the customer's dial tone service, whatever its precise definition, can be obtained by the dial tone tester method.

Recently revised tables for the capacity of step-by-step line finders have been published for Bell System use based on Palm's formula using a factor of $j = 5$. This was selected as being slightly conservative for

most applications after reviewing the above Sterling-3 results and other line finder data collected in step-by-step offices.

MEASUREMENT OF THE j FACTOR BY CLASSES OF SERVICE

As indicated previously, the data recorded on the tapes showed the states of being busy or idle and of changes in these states for line finders and the associated trip circuits. A fully equipped line finder group of 400 lines has ten trip circuits each of which serves two sub-groups of twenty subscriber lines in the following manner. When a line originates a call its line relay is operated. This causes a ground to appear on a lead which is common to all twenty line relays in the sub-group and starts a line finder hunting for the calling subscriber's line. As soon as this hunt is completed the cutoff relay associated with the calling line operates and disconnects the line relay, removing the ground (unless, of course, another line in the sub-group has originated a call in the meantime). During periods of overload when line finders are not immediately available, the ground due to a single subscriber will persist until:

1. A line finder is obtained, or
2. The subscriber abandons the attempt, or
3. The subscriber receives an incoming call which operates the cut-off relay.

The twenty leads from the trip circuit sub-groups were brought out to the pen recorder and a record taken of the grounds that occurred on each lead. Except for the possibility that more than one subscriber is waiting for service at the same time on a given trip circuit sub-group, the record of the occurrences of the grounds gives a substantially accurate⁷ record of the demands for service and of the number of calls waiting for service. Hence an analysis of the events occurring on the trip circuit sub-groups and on the line finders as recorded on the tapes gives a means for determining H . The quantity H was introduced in equation (2) in the term

$$\frac{(x + 1 - c)}{H} f(x + 1) \quad (18)$$

For convenience in the ensuing discussion this term will be replaced by

⁷ To obtain absolute accuracy would require the use of a pen recorder with one pen for each of the 400 subscribers served on a line finder group plus one for each line finder.

the equivalent expression:

$$N_y = \frac{y}{H} f(z + y) \quad (19)$$

where N_y = The average number of waiting calls that drop out per unit of time during the state $(z + y)$.

y = The number of waiting calls.

z = The number of line finders occupied with calls.

$1/H$ = A measure of the rate at which calls tend to abandon waiting.

$f(z + y)$ = The proportion of time that the state $(z + y)$ exists.

On the tapes we can measure $f(z + y)$ and count N_y . Hence H can be determined. The result is a statistical quantity subject to many chance factors. In the actual analysis of a tape, the composite average value of H was determined for all possible observed states where calls were waiting. By an analogous process, the composite average value of h for all possible observed states where calls were being served by line finders was determined. Also as a side computation, a composite average value of h' for calls that were served by line finders but for which no peg counts were scored was determined. This value of h' is included in h on the basis that data for engineering line finders consist of estimated calls based on peg counts and of holding times which include an allowance for these short holding time calls. The average values of H , h and of the j factor for the four classes of service studied are given in Table III.

The results for H by individual half hours and by various percentages of dial tone delays greater than three seconds are shown on Figs. 7(a) to 7(d) respectively for the four classes of service. On some of these figures an upward bulge may be noted in the center. This is not considered to be characteristic of the habits of the subscribers but is the overall effect resulting from a number of arbitrary rules followed in making the analysis in order to simplify the work and to offset par-

TABLE III

	Average Values in Seconds		$j = h/H$
	H	h	
Message rate individual.....	24	159	6.6
Message rate two-party.....	42	243	5.8
Flat rate individual.....	27	176	6.5
Coin.....	74	153	2.1

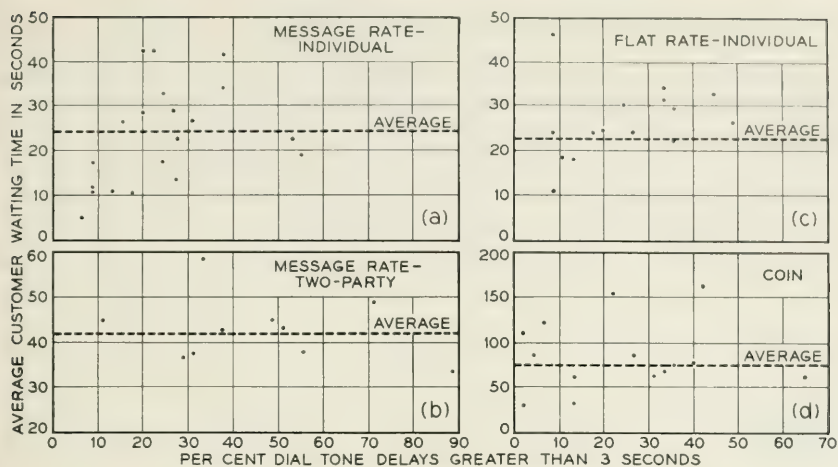


Fig. 7—Average customer waiting time (H).

tially the effect of occasionally having two or more calls waiting on one trip circuit sub-group. The rules and the reasons for them will be described with the aid of Fig. 8, which shows a hypothetical section of one of the tapes. The rules were as follows:

1. Initial Overlap

Referring to Fig. 8, at t_1 a subscriber has initiated a request for service. At t_2 a line finder rises to serve the subscriber. At t_3 the subscriber receives service. This case is typical of a subscriber receiving prompt dial tone service.

The span from t_1 to t_2 was difficult to measure accurately because, for the usual case, it was about the same as the maximum error due to misalignment of the recorder pens. It was not measured unless the combined span from t_1 to t_3 exceeded one second.

The span from t_2 to t_3 involves an overlap, it represents a period when a line finder is busy hunting for the terminal of the subscriber who originated the request for service. It also represents a period when a subscriber is waiting for service. In the analysis this span was treated as a case where a line finder was busy with a call and not as a call waiting for service.

If the span from t_2 to t_3 and all similar cases had been treated as calls waiting for service and if in addition all spans from t_1 to t_2 which were not measured had also been treated as calls waiting for service, the average values for H would have increased slightly for each class of service.

2. Three Second Rule for the Bridging of Calls

Referring to Fig. 8, again, at t_5 a request for service is originated on trip circuit 5 and at t_7 this request is withdrawn. At t_{10} apparently a new request for service is initiated which is then withdrawn at t_{11} . From manual service observations it is found that subscribers often flash when dial tone is slow. A few pens were used to observe individual subscribers, and Fig. 9 shows a case where a subscriber made several flashes when his tone was slow. When a subscriber flashes it appears as though he

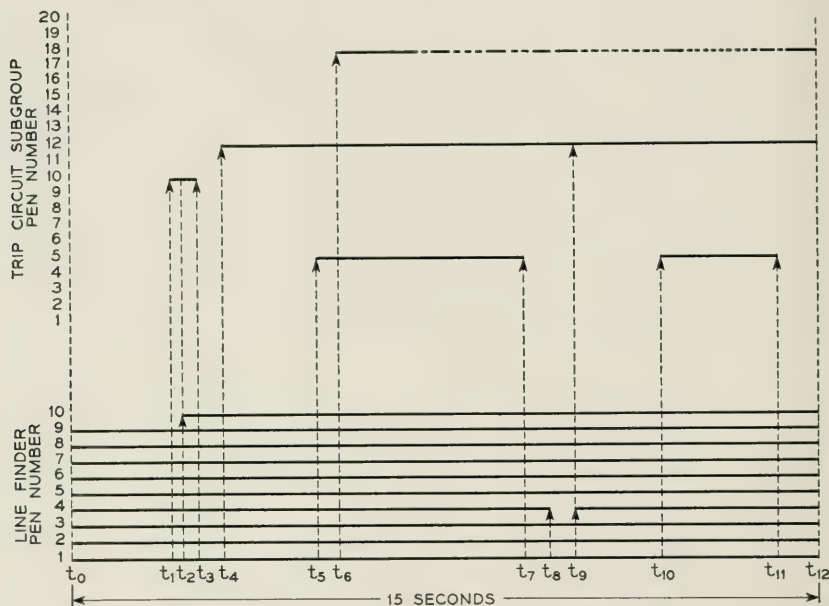


Fig. 8—Section of a hypothetical tape showing activities on trip circuit sub-groups and on line finders.

were making several bids for service. Actually he is making only one real bid. On the trip circuit sub-group pens it was generally impossible to distinguish between flashes and requests for service by two or more subscribers. To resolve this problem many observations of subscriber lines recorded on the tapes were examined from which it was concluded that no great error would result if a break in the demand for service on a trip circuit sub-group of less than three seconds were considered as a flash and was to be bridged, and a break greater than three seconds was to be considered as the termination of one call attempt and the start of another

3. Treatment of Cases Where Two or More Calls Were Found to be Waiting on One Trip Circuit Sub-Group

The occurrence of several calls waiting on one trip circuit was occasionally noted in the analysis. Referring to Fig. 8, a case is shown on trip circuit sub-group 12. At t_9 a line finder is seized. Trip circuit sub-group 18 shows that a subscriber is dialing before tone. The appearance of dial pulses on this trip circuit indicates that only one subscriber is demanding service otherwise the dialing would not show. Trip circuit sub-group 12 however appears to have two or more requests for service. One of these requests for service began at t_4 . The start of the second request occurred somewhere between t_4 and t_9 , perhaps half-way between. At t_9 , one of the requests was served by a line finder. To simplify the handling of such cases, the assumption was made that the first attempt started at t_4 and ended at t_9 and the second request started

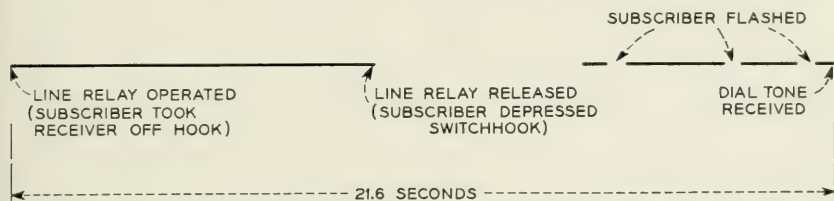


Fig. 9—Example of a customer flashing for dial tone (Tape made October 18).

at t_9 . The effect of this is to understate by an indeterminate amount the average value of H for each class of service. This understatement should be noticeable for the higher degrees of overload because the occurrence of several calls on one trip circuit sub-group is then most likely to occur.

The effect of several calls simultaneously waiting in a trip circuit sub-group and of one or more calls dropping out is to overstate the magnitude of H . For instance if two simultaneous call attempts of five seconds each overlap for one second and both attempts are abandoned, the apparent average waiting period is nine seconds, whereas it should be five seconds. It is believed that the above three rules tend to create understatements which roughly balance this type of overstatement.

DISTRIBUTION CURVES OF SIMULTANEOUS CALLS

The detailed analysis of the tapes provided distributions of simultaneous calls. For each class of service studied these distributions can be compared with theoretical distributions derived from the generalized trunking formula using the j factors developed in the analysis. Several

such comparisons are shown on Figs. 10 and 11. The agreement is quite good in most cases.

SUBSCRIBER DIALING HABITS AS OBSERVED WITH A MONITORING CIRCUIT ON A SENDER WITH INDUCED DIAL TONE DELAYS⁸

As a separate study a series of tests was made by means of a monitoring circuit on one of the senders serving in common the subscribers in the Sterling-3 and Main 2 central offices, for the purpose of obtaining further information on subscriber dialing characteristics under overload conditions. A large amount of data was collected on the time intervals from the seizure of the sender to the first action taken by subscribers when encountering dial tone delays, the latter being introduced under the control of the observer.

The monitoring circuit was wired to a particular sender in a group of 100 serving all classes of subscribers. When the circuit was in use, the only irregularity introduced was that the dial tone could be delayed even though the sender was actually available to the subscriber. The delay did not affect the sender in its functions if the subscriber elected to dial before tone.

The sender monitoring circuit provided the following four features:

1. A receiver was bridged across the tip and ring leads in the sender so that an observer could hear certain actions taken by a subscriber connected to the sender. The sender was of course disconnected before conversation.

2. The observer was able to preselect one of several intervals by which dial tone was delayed on successive calls served by the sender. This was accomplished with a capacitance-resistance-vacuum tube circuit.

3. By means of a timer which started when the sender was seized, the observer was enabled to note elapsed time intervals to the occurrence of the various actions of the subscribers. The reading of the time of the first action of a subscriber had to be made when the second hand was in motion, which introduced certain errors later to be discussed.

4. By means of colored lamps the observer was able to classify all calls observed as being message rate, flat rate or coin.

During the sender dial tone delay tests, observations were made only during the afternoons when the flow of traffic was light and the probability of a subscriber obtaining a delay before reaching the sender was a minimum.

⁸ Based on an unpublished report by W. A. Reenstra.

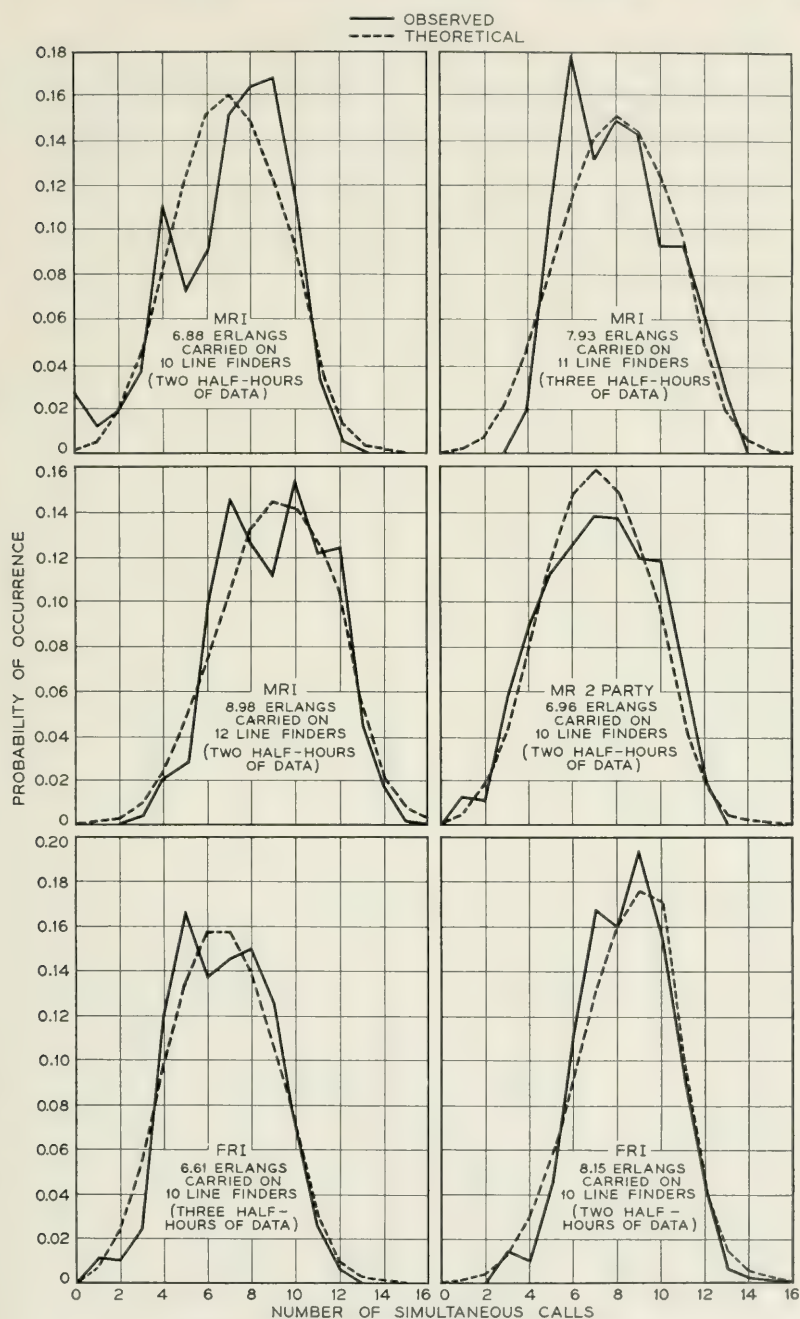


Fig. 10—Distributions of simultaneous calls.

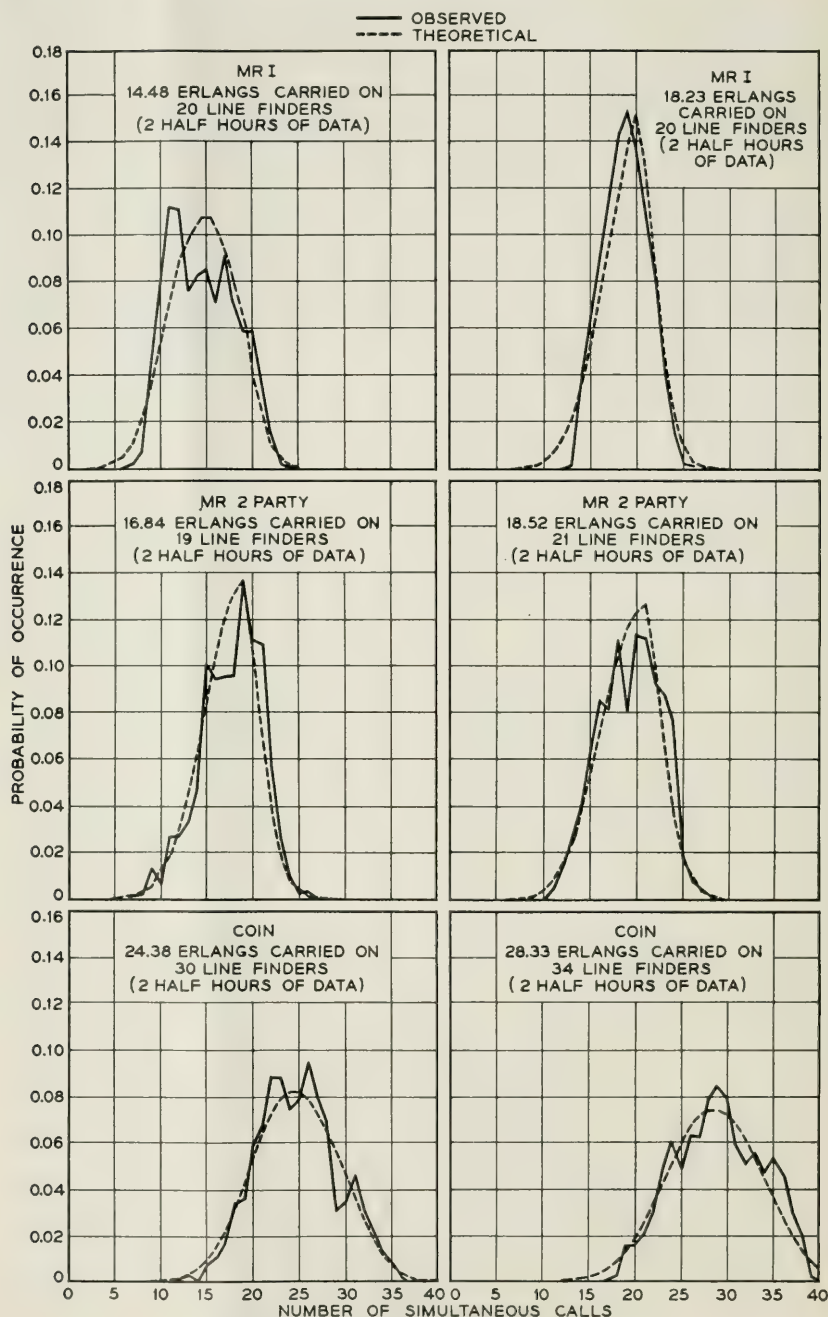


Fig. 11—Distributions of simultaneous calls.

The observer was provided with a means for introducing either no delay or one of four values of delay 2, 5, 10 or 15 seconds into the sender dial tone circuit. The observer took 50 observations using a particular value of dial tone delay and then shifted to another so that no particular value of delay would become evident to the customers during an afternoon's test. Each group of 50 observations comprised a mixture of message rate, coin and flat rate calls in the approximate proportions of 13 to 6 to 1, representing the respective volumes of traffic from these classes of service during the afternoon periods. It was not possible to distinguish PBX lines or two-party lines from the bulk of the message rate data nor PBX lines in the flat rate data, although to a limited extent the observer could identify PBX dialing by the generally faster pulsing. The coin data represent both public and semi-public customers.

Fig. 12 is a diagram for explaining the results shown on Figs. 13, 14 and 15 for the message rate, flat rate and coin classes of service, respectively as obtained with the sender monitoring circuit. Fig. 12 was obtained by the application of fitting curves to those message rate data of Fig. 13 for which a dial tone delay of five seconds was introduced by the observer. In the interval from $t = 0$ to $t = 5$ seconds, three curves A, B and C represent the per cent of subscribers still waiting at time t for dial tone. Curve A and its extension beyond $t = 5$ seconds represents the action of subscribers who would dial their calls before tone if dial tone were delayed indefinitely. Curve B and its extension beyond $t = 5$

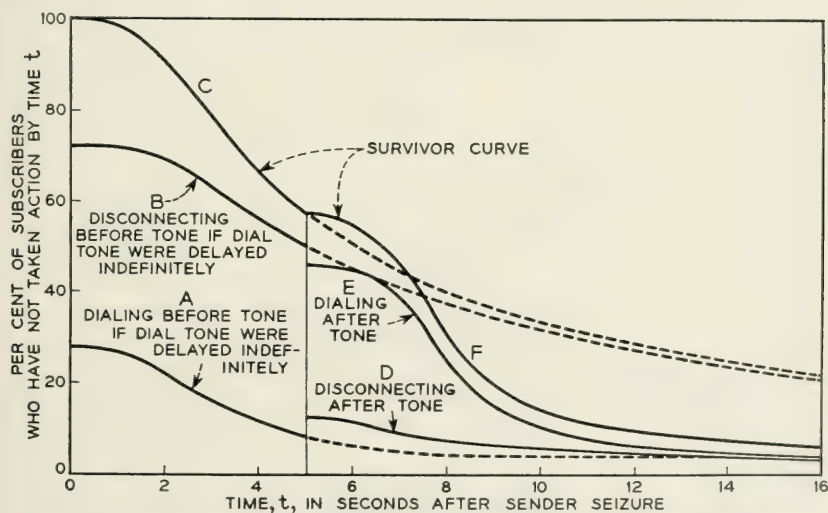


Fig. 12—Explanatory chart for sender monitoring observations; dial tone at $T = 5$ seconds.

seconds represents actions of subscribers who would disconnect if dial tone were delayed indefinitely. Curve C and its extension is the sum of the other two. Curve D in the region beyond $t = 5$ seconds represents the actions of subscribers who disconnect after tone, curve E represents the actions of subscribers who will dial their calls after tone and curve F represents the sum of the lower curves. Of interest is the fact that for an interval of about two seconds following dial tone (at $t = 5$ seconds), the observed total survivor curve F lies above the extended portion of

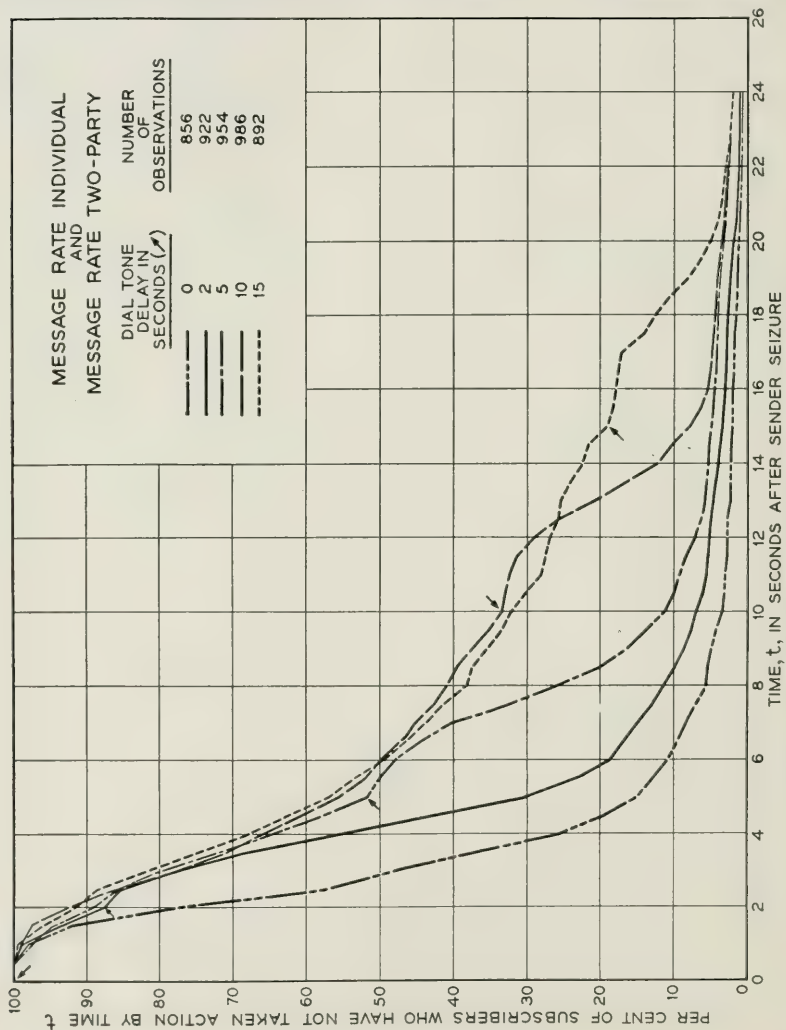


Fig. 13—Results of sender monitoring observations.

the hypothetical survivor curve C for infinitely delayed dial tone. This indicates that most of the subscribers who would have abandoned their attempts during this interval abruptly changed their minds and then consumed a noticeable interval of time after hearing tone before starting to dial. Thus, as might be anticipated, the subscribers exhibit a reaction time.

Fig. 13 shows the results in terms of survivor curves that were observed for the message rate class of service. Five sets of curves are

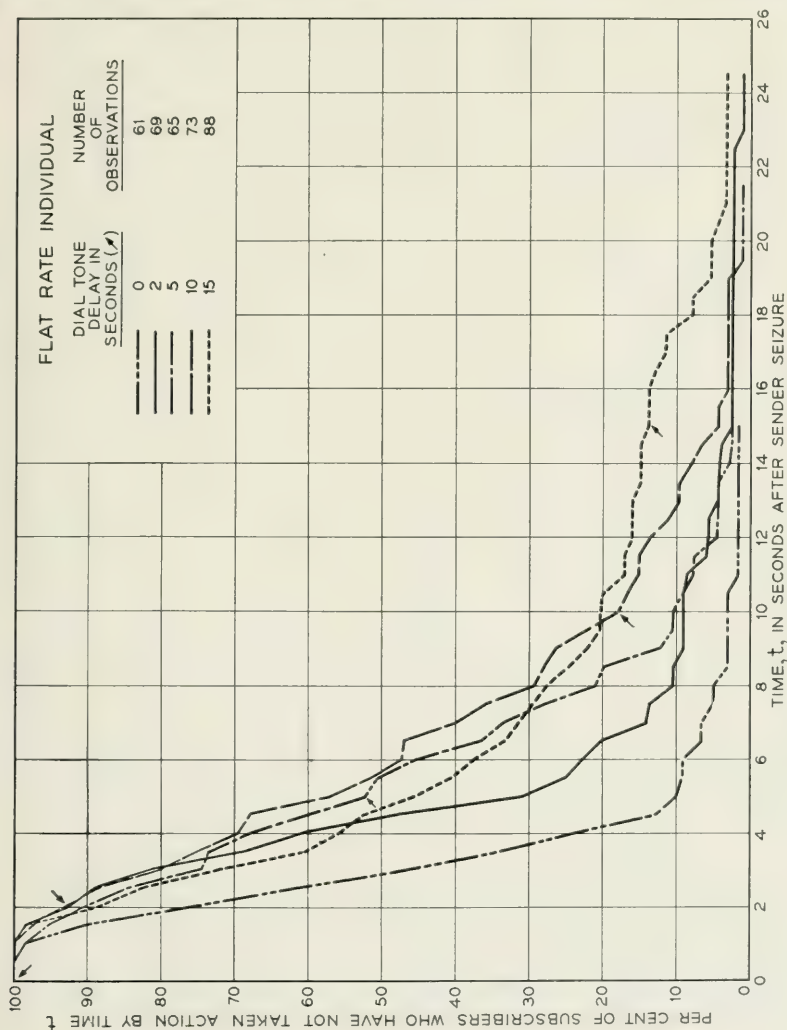


Fig. 14—Results of sender monitoring observations.

presented, namely for 0, 2, 5, 10 and 15 seconds of dial tone delays from the instant of sender seizure. It should be noted that the general contour of the various curves up to the receipt of dial tone and when extended beyond gives an estimate of the survivor curve for dial tone delayed indefinitely. Fig. 14 shows the results for the flat rate class of service (the data here are relatively meager), and Fig. 15 shows the results for coin customers.

Fig. 16 indicates for the three classes of service the progressive changes,

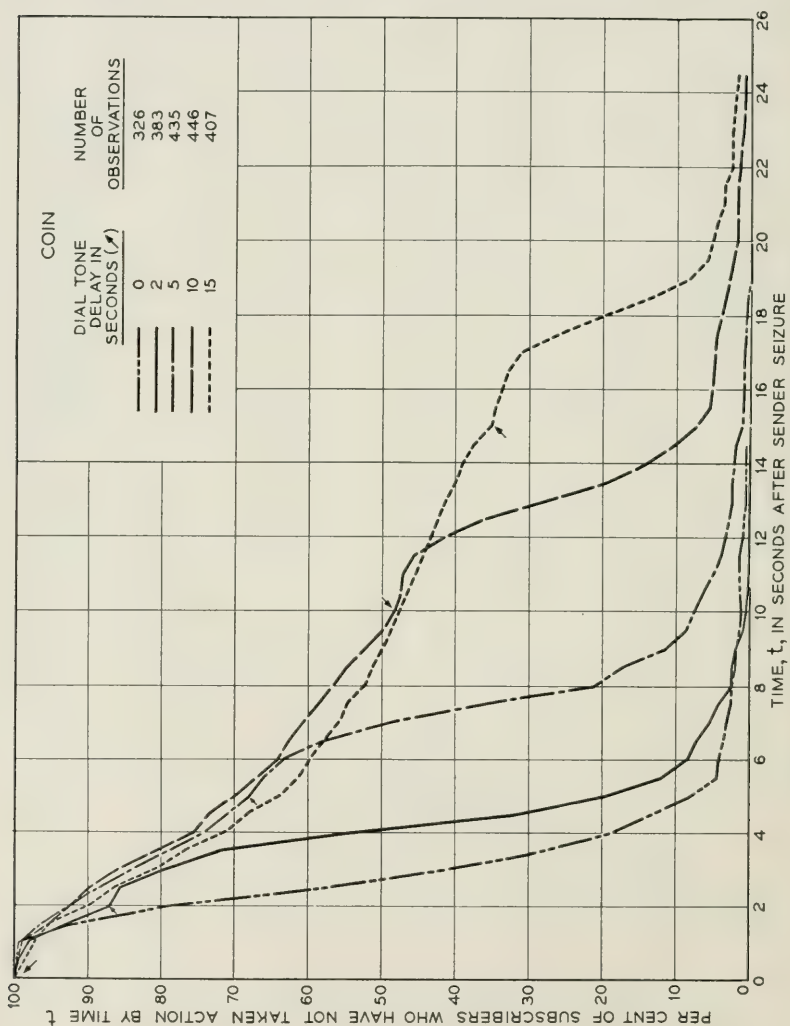


Fig. 15—Results of sender monitoring observations.

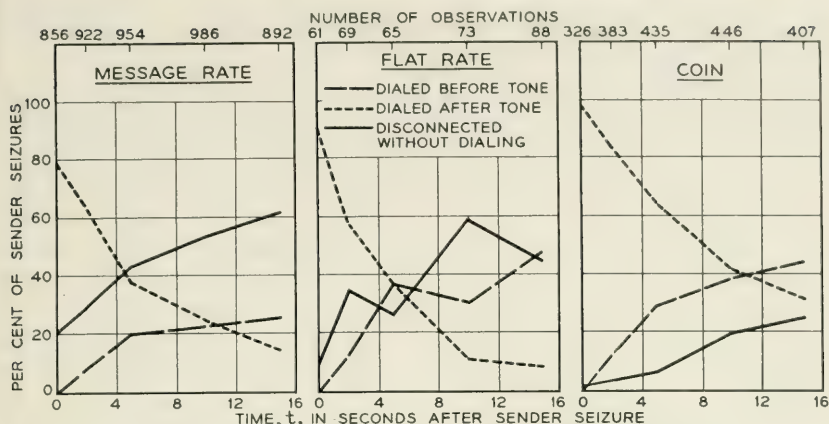


Fig. 16—Results of sender monitoring observations.

with increasing dial tone delay, of the per cent sender seizures resulting in dialing before tone, in dialing after tone and in disconnections for the five dial tone delay intervals studied.

Some general comments concerning Figs. 13 to 16 may be made.

1. There is a striking contrast between the message rate and coin classes of service. This may be due to the immediate financial stake that a coin customer has in his call. He is reluctant to disconnect before dial tone.

2. The results for the message rate and flat rate classes of service appear to be similar at the shorter dial tone delays; at the longer delays a higher proportion of flat rate subscribers have already dialed before dial tone. This apparent discrepancy may be due to the relatively small number of flat rate observations. Flat rate service in Sterling-3 and Main-2 was principally for professional people, such as doctors and nurses, who were thought to be more demanding than ordinary subscribers for prompt dial tone service. These results therefore should not be considered characteristic of flat rate customers generally.

3. Detailed analysis of the data (results not presented in this article) indicated that the distributions of time to first pulse for subscribers not observing tone, and for those waiting for tone, are quite similar for the three classes of service.

4. Only the unsmoothed raw data have been shown on Figs. 13 to 15 since certain inadequacies were detected in the observations. These were due to observer reactions and reading errors discovered as a result of comparing preliminary practice sender monitor test results with a simultaneous record obtained by the 100 pen tape recorder. The overall

effect is that the data obtained by the observer are generally displaced outward along the time axis by about 0.8 second.

5. The message rate data were for individual, PBX and two-party subscribers and the flat rate data were for individual and PBX subscribers. Furthermore, certain of the flat rate subscribers had auxiliary message rate service. It seems likely that different characteristics would be obtained for the individual, PBX and two-party subscribers since there appear to be reasons for expecting significant differences in their dialing habits. The PBX operator is in a position to "shop" for telephone service. If she fails to get dial tone on one outgoing line, she can try any other free line. This can also be done by subscribers with multi-line service. This "shopping" for service tends to produce a large volume of disconnections when dial tone is slow. The individual and the two-party subscribers cannot do this and hence they can be expected to show fewer disconnections.

6. The results are in terms of intervals of time from the instant the sender is seized. It would, of course, be preferable to have these results in terms of time from receiver off hook. Since on the average the sender is seized in a time interval of about the same magnitude as that of the reaction time of the observer, Figs. 13 to 15 can be read approximately correctly when the abscissas are redesignated "time in seconds from receiver off hook." The foregoing results have been presented to furnish an increased understanding of subscriber dialing habits. In the next section additional results based on individual line records taken on the tapes are presented.

SUBSCRIBER DIALING HABITS OBSERVED BY INDIVIDUAL LINE RECORDS

As indicated in the previous section, the results obtained by means of the sender monitor tests were subject to certain shortcomings, hence data taken on individual lines with the 100-pen recorder have been analyzed to augment the information concerning the dialing habits of subscribers.

As noted heretofore, several of the pens on the 100-pen recorder were available for taking observations on subscribers lines. Two pens were used per subscriber line, one recorded the operation of the subscriber's line relay while the other pen marked whenever the subscriber's line was busy. On an originating call both pens started marking when the subscriber initiated a call. When a line finder was obtained, the line relay pen ceased marking and it was presumed that the subscriber obtained dial tone at that instant. Dialing, hang-up and flashing by a

subscriber after receipt of dial tone are noted by breaks in the markings of the line busy pen. Dialing, hang-up or flashing before dial tone are noted by simultaneous breaks in the markings of both pens. Various intervals can be measured and the call attempts classified accordingly.

Observations were obtained in the foregoing manner during the course of the Sterling-3 line finder tests on the following numbers of subscriber lines:

Message rate—residential.....	87
—business.....	23
—PBX.....	12
—two party.....	32
Flat rate individual.....	7
Coin.....	21
	<hr/>
	182

The observed data were classified for each of the above six types of subscribers in terms of the following categories:

1. Time to subscriber action before receipt of dial tone.
 - a. Time from receiver off hook⁹ to first digit dialed by subscriber.
 - b. Time from receiver off hook to disconnecting action by the subscriber.
2. Time from receiver off hook to receipt of dial tone.
3. Time to subscriber action after receipt of dial tone.
 - a. Time from receipt of dial tone to first digit dialed by subscriber.
 - b. Time from receipt of dial tone to disconnecting action by the subscriber.

Because the data developed in this section are compared with both the *j* factor analysis and the sender monitor test results and because the treatment of subscriber dialing habits before dial tone for each of these items is different, categories 1a and 1b are analyzed in two ways. In the *j* factor analysis, all actions of a subscriber prior to dial tone, except a disconnect, were considered to be one continuing demand for service. Thus for the first analysis, cases of dialing, flashing and short disconnections before tone lasting less than three seconds were ignored. In the sender monitor tests the only items considered were the time to the first digit dialed by a subscriber and the time, if no dialing occurred, to the release of the sender by the subscriber. Thus for the second analysis, cases of dialing before tone and flashing or disconnections before tone

⁹ When a customer initiates a call, the line relay operates. For individual line and two-party subscribers this occurs when the subscriber takes his receiver off the hook. For coin customers this occurs when the customer has taken his receiver off the hook and made a proper deposit. For a PBX line this occurs when the PBX attendant has established a connection to an outside line. All of this is collectively termed "receiver off hook."

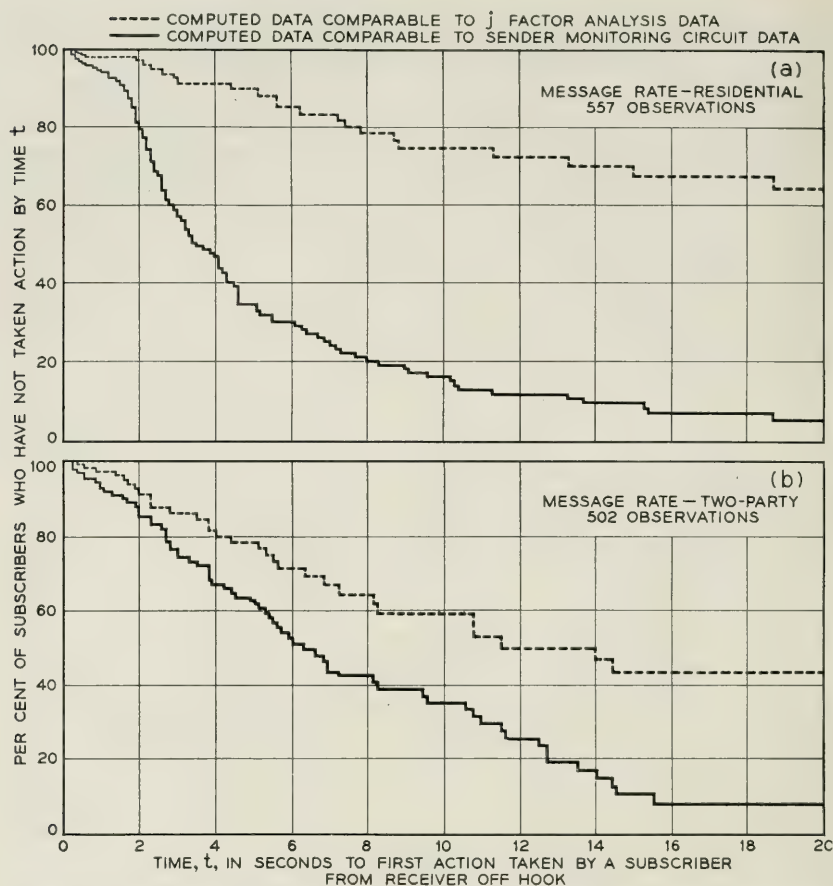


Fig. 17—Subscriber dialing habits based on individual line records when dial tone is delayed indefinitely.

that caused the sender to release were each counted as separate attempts. The results arrived at are shown as survivor curves on Figs. 17(a) to 19(b) for each of the six types of subscribers studied.

The survivor curves were developed by considering events during 0.1 second intervals. The number of cases of dialing before tone and the number of cases of disconnection before tone occurring during a 0.1 second interval were divided by the number of cases waiting for dial tone at the start of the interval. This ratio was considered to be a retirement rate. The complement of this rate gave a survival rate. By a progressive multiplication of survival rates from time, $t = 0$, the resulting survivor curves were obtained. Those cases receiving dial tone

during a particular 0.1 second interval are omitted from the number of cases waiting for dial tone at the start of the next interval.

In the development of the generalized trunking formula, the assumption was made that the waiting times of calls infinitely delayed have an exponential distribution. By assuming that the plots for the survivor curves in the development comparable to the j factor analysis are exponential distributions, it is possible by reading the value of t corresponding to 36.8 per cent of the subscribers still waiting for dial tone to obtain estimates of the values of H for the six types of subscribers. These estimates, most of which were obtained by extrapolation, are compared in

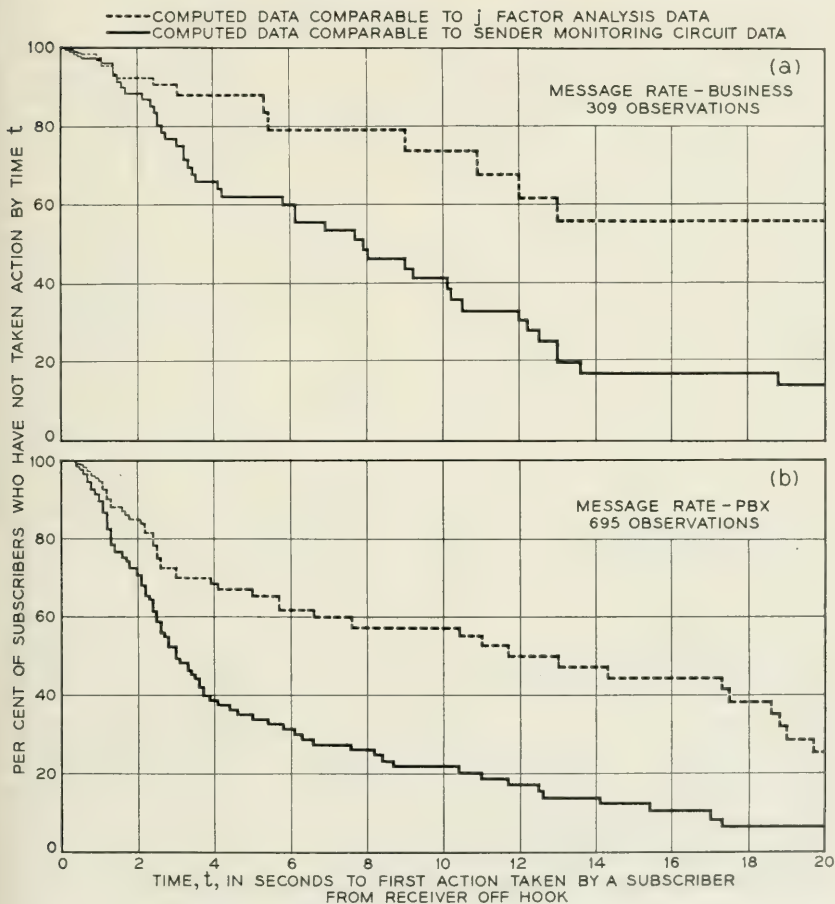


Fig. 18—Subscriber dialing habits based on individual line records when dial tone is delayed indefinitely.

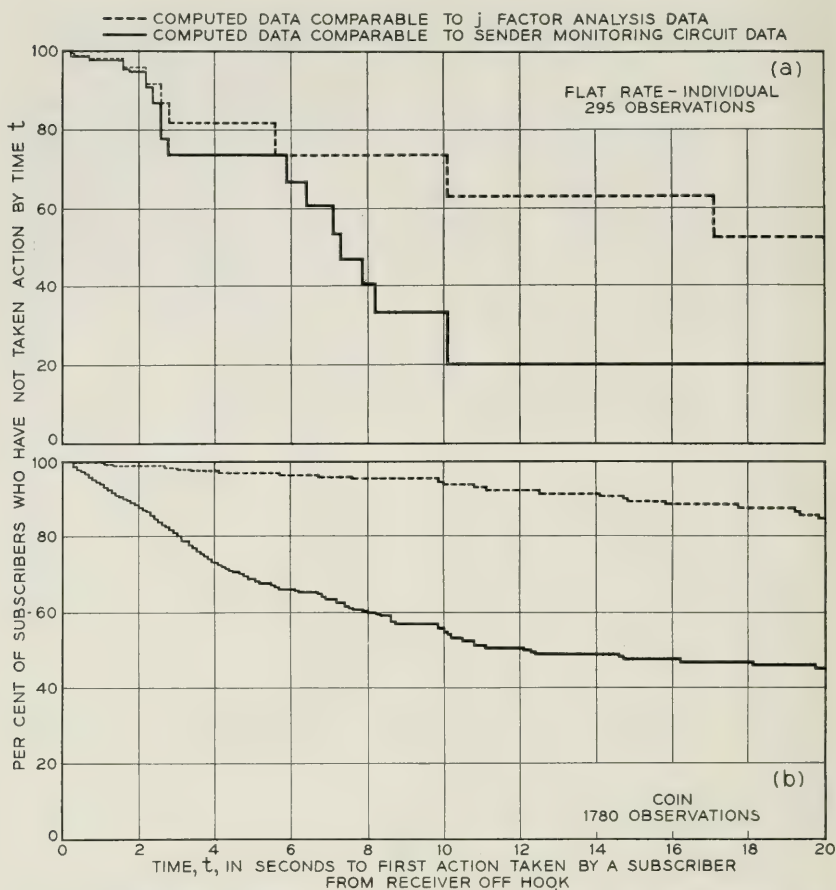


Fig. 19—Subscriber dialing habits based on individual line records when dial tone is delayed indefinitely.

Table IV with the values developed earlier in connection with the j factor analysis.

These results are not considered to be inconsistent since the tails of the survivor curves were constructed from very meager data. Conclusions based on Figs. 17(a) to 19(b) should therefore be regarded as having qualitative value only. The results principally indicate that the waiting-for-dial-tone characteristics of subscribers clearly vary with the different classes of service.

Comparisons between the lower survivor curves on Figs. 17(a) to 19(b) and the curves on Figs. 13 to 15 of the sender monitor tests are indicated by the percentages given in Table V of subscribers waiting for dial tone 5, 10 and 15 seconds from the time they requested service.

TABLE IV

	Estimated Values of H in Seconds	
	From figures 17(a) to 19(b)	From j factor analysis
Message rate—residential.....	45*	} 24
— business.....	33*	
— PBX.....	19	
— two-party.....	24*	
Flat rate individual.....	32*	42
Coin.....	110*	27
		74

* Rough extrapolated values

TABLE V

	Percentages of Subscribers Waiting for Dial Tone					
	Service Observation Figs. 17(a) to 19(b)			Sender Monitor Tests Figs. 13 to 15		
	5 secs	10 secs	15 secs	5 secs	10 secs	15 secs
Message rate						
Residential.....	34%	16%	10%	} 55%	} 33%	} 19%
Business.....	62	41	16			
PBX.....	35	22	13			
Two party.....	62	35	11			
Flat rate individual.....	74	34	20	55	19	14
Coin.....	68	55	47	67	48	35

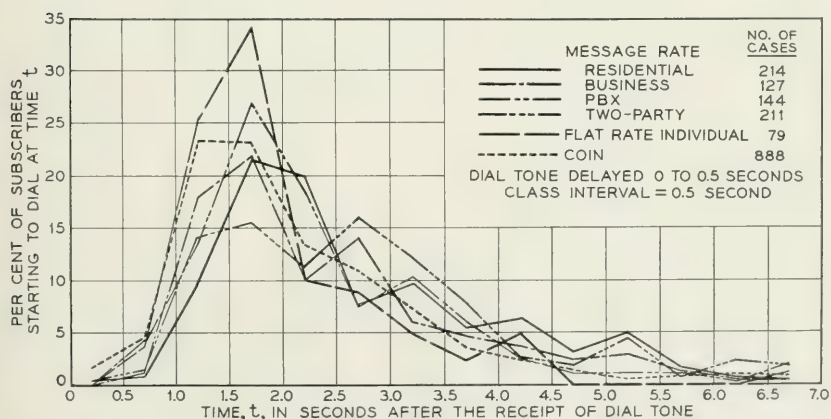


Fig. 20—Distributions of the start-to-dial times of subscribers who dial after the receipt of dial tone.

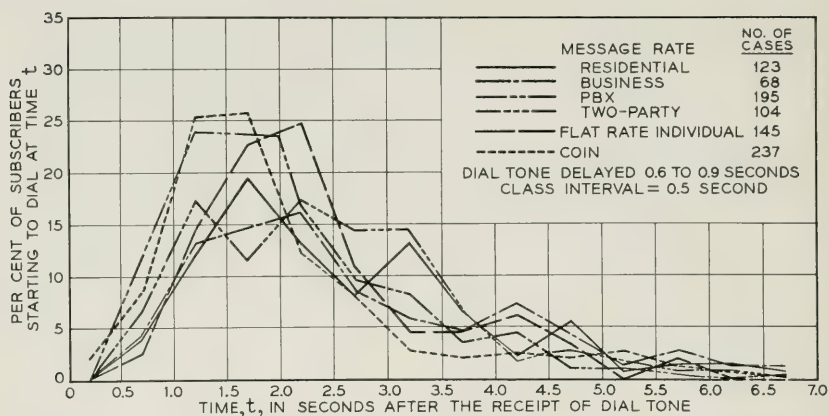


Fig. 21—Distributions of the start-to-dial times of subscribers who dial after the receipt of dial tone.

These results agree reasonably well when it is recalled that parts of the individual line data were scanty and that the sender monitor tests included the effects of observer reactions.

Once dial tone is received, it appears that all types of subscribers tend to follow a uniform dialing pattern. Figs. 20 to 23 show for a class interval of 0.5 second the distributions of the per cent of subscribers who dial at time t for the six types of subscribers studied. Figs. 20, 21 and 22 show the distributions when dial tone is received from 0.0 to 0.5, 0.6 to 0.9 and 1.0 to 1.9 seconds after dial tone, respectively. These curves

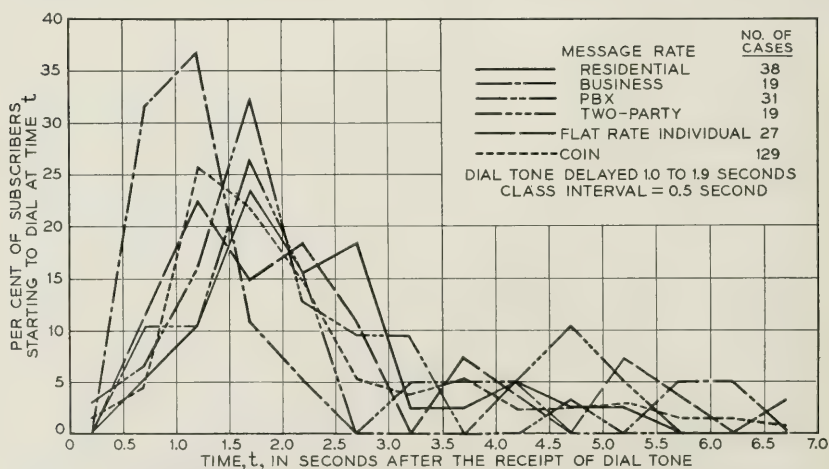


Fig. 22—Distributions of the start-to-dial times of subscribers who dial after the receipt of dial tone.

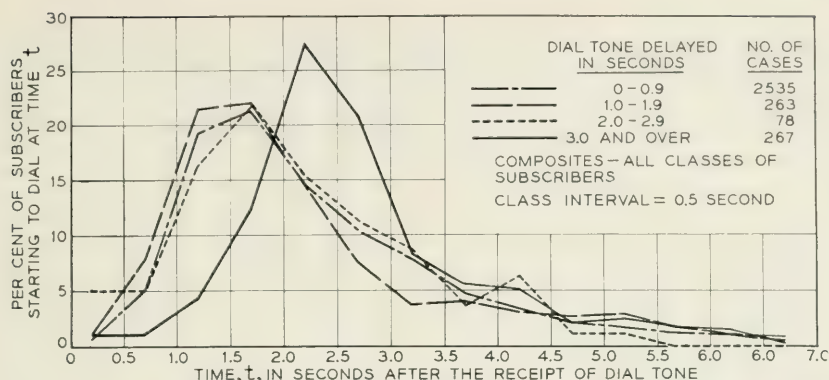


Fig. 23—Distributions of the start-to-dial times of subscribers who dial after the receipt of dial tone.

indicate that a strong similarity exists among the six types of subscribers with regard to their dialing patterns once dial tone is received.

Fig. 23 shows composite dialing distributions of the six types of subscribers for four dial tone delay intervals. For dial tone delays less than 3 seconds the dialing patterns are seen to be similar. Beyond 3 seconds delay the start-to-dial curve shifts outward, although the data are too scattered to indicate closely where the movement begins.

CONCLUSION

The foregoing report of the results of the tests conducted at the Sterling-3 central office indicates that the dialing habits of subscribers waiting for dial tone can be observed and analyzed to develop so-called j factors for use in a more general trunking formula than has been employed until recently in the Bell System. The report also presents descriptive data regarding the patterns subscribers follow when waiting for dial tone.

ACKNOWLEDGEMENT

The writers gratefully acknowledge the help given by the personnel of the Long Island Area of the New York Telephone Company, by their associates and others in the various phases of the study and particularly by C. F. Bischoff who had a major part in conducting the tests.

Selective Fading of Microwaves

BY A. B. CRAWFORD AND W. C. JAKES, JR.

(Manuscript received October 25, 1951)

The results of an extended survey of microwave propagation over two line-of-sight paths in New Jersey are described. Angle-of-arrival measurements at 1.25-cm wavelength and selective fading observations in a 450-mc frequency band centered at 3950-mc show that the severe fading can be explained in terms of multiple-path transmission. A computer of the analogue type was built to simulate the more complicated selective fading patterns.

INTRODUCTION

During the past few years, studies of microwave propagation have been made by the Radio Research group at the Holmdel Laboratory over two paths located in eastern New Jersey. Both of these are line-of-sight paths which might be considered to be typical links in a cross-country microwave radio relay circuit.

In conducting these studies, the usual continuous recordings of signal levels were made but the greater interest was centered in special experiments designed to reveal more of the processes which can cause fading. The most relevant information has been obtained by exploring the incident wavefronts with a narrow-beam scanning antenna (angle-of-arrival studies) and, more recently, by observing the transmission characteristics of the paths by means of a frequency-sweep technique and also by the use of very short pulses.

Some results of angle-of-arrival observations have been reported previously¹ and a companion paper describes the transmission tests conducted with very short pulses.² The present paper describes some of the observed mechanisms associated with fading, presents typical data obtained with the narrow-beam scanning antenna and gives examples of the frequency-sweep observations, illustrating the frequency selective

¹ W. M. Sharpless, "Measurement of the Angle of Arrival of Microwaves," *Proc. I.R.E.*, **34**, Nov. 1946, pp. 837-845. A. B. Crawford and W. M. Sharpless, "Further Observations of the Angle of Arrival of Microwaves," *Proc. I.R.E.*, **34**, Nov. 1946, pp. 845-848. H. T. Friis, "Microwave Repeater Research," *Bell System Tech. J.*, **27**, Part I, "Propagation Studies" by A. B. Crawford, Apr. 1948, pp. 183-246.

² O. E. DeLange, "Propagation Studies at Microwave Frequencies by Means of Very Short Pulses," *Bell System Tech. J.*, **31**, Jan. 1952, pp. 91-193.

nature of the fading. Some data derived from the continuous recordings of signal levels are presented in an appendix.

The angle-of-arrival observations were made at a frequency of 24,000 megacycles. The frequency-sweep experiment and the recordings of signal levels were made in the 3700 to 4200 megacycle frequency band as were the short pulse observations described in the companion paper.

GENERAL DISCUSSION OF PROPAGATION PHENOMENA

The map of Fig. 1 shows the location of the experimental transmission paths. The path between Crawford Hill and Southard Hill is 17 miles long and clears the intervening terrain by 65 feet, approximately one Fresnel zone at a frequency of 4000 megacycles. The other path, between Crawford Hill and a 100-foot tower on the Murray Hill Laboratory property, is 22.8 miles long and has clearance of 280 feet. Fig. 2 shows the profiles of these two paths.

The general characteristics of over-land microwave transmission are well known and need be reviewed only briefly. During the daytime hours, when the lower atmosphere is thoroughly mixed by rising convection currents and winds, the signals are normally stable and are near

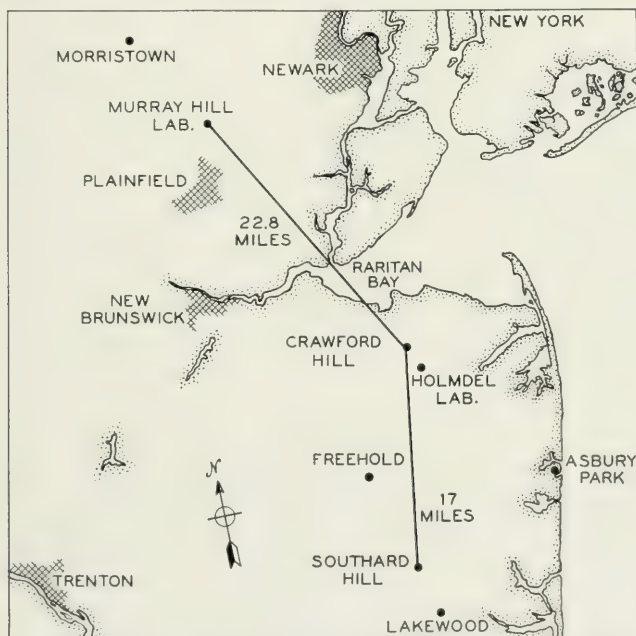


Fig. 1—Map showing location of the experimental transmission path.

the free space levels. Also during the winter months, when the humidity content of the atmosphere is low, signal variations are usually very small. However, on clear summer nights with little or no wind, non-uniform distributions of temperature and humidity can create steep dielectric constant gradients in the lower atmosphere, thus causing anomalous propagation and fading.

When fading occurred on our experimental transmission paths, an alarm circuit connected into the continuously recording equipment was arranged to operate when the signal level dropped below a predetermined value. This enabled observers to be present during severe fading periods

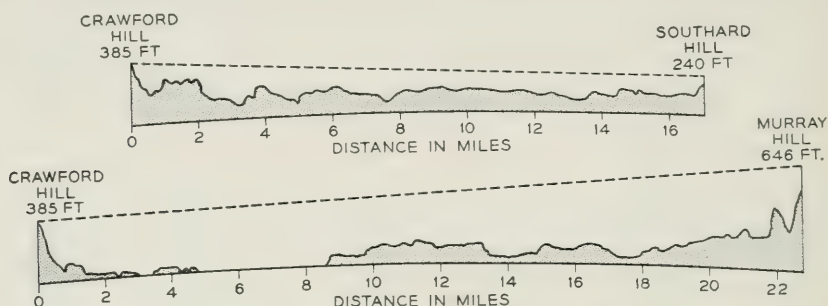


Fig. 2—Profiles of the transmission paths.

and to seek, by means of the special experiments, to determine the causes of the fading.

Although it has not been possible to provide satisfactory explanations for all of the observed fading phenomena, much of the fading (occasions when the signals are depressed to levels 15 to 20 decibels or more below the normal daytime value) can be explained qualitatively in terms of simple ray pictures. Fig. 3 is intended to illustrate some of the observed fading phenomena. The case of multiple path transmission, the most common cause of fading on either transmission path, is shown in Fig. 3(a). Two, three and sometimes more signal components are found to arrive at various angles in the vertical plane, usually above the line of sight. Wave interference among these components produces fading, the severity of which depends upon the relative amplitudes and delays of the components. At these times, different frequencies fade differently and the signals received on two vertically spaced antennas also fade differently. The use of either frequency or space diversity would be effective in this type of fading.

A relatively rare type of fading, observed only on the Murray Hill path, is believed to be caused by the mechanism illustrated in Fig. 3(b).

Here a reflecting layer is situated between the heights of the transmitter and receiver. The signal then suffers attenuation due to reflection of part of the energy from the direct path. Widely separated frequencies are affected in like fashion and the outputs of antennas spaced for diversity reception tend to be in agreement although the fine structure fading is usually different.

On neither of the experimental transmission paths is there a regular ground-reflected component of any consequence. Due to the roughness of the ground and the presence of vegetation, the effective reflection coefficient is of the order of 0.2 for either path. Ground reflections thus play

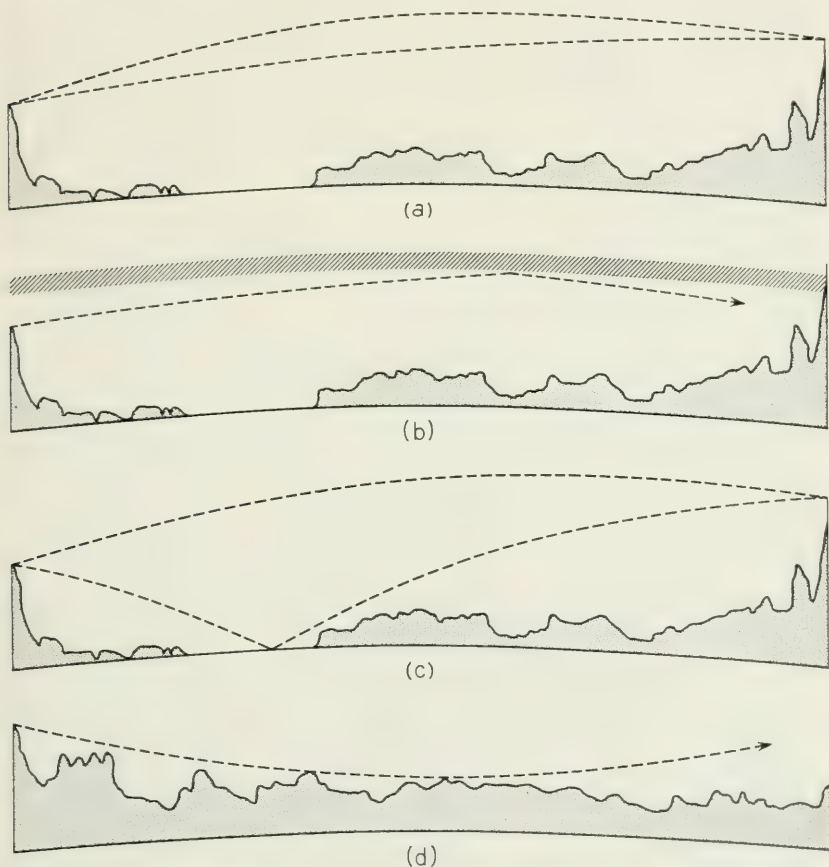


Fig. 3—Possible ray paths involved in severe fading. (a) Multiple path transmission. (b) Attenuation by reflection from an elevated layer. (c) Abnormal water reflection on the Murray Hill-Crawford Hill path. (d) Substandard conditions on the Southard Hill-Crawford Hill path.

no significant part in the fading picture with the exception of the situation illustrated in Fig. 3(c). Occasionally on the Murray Hill path, conditions of atmospheric refraction are such that a strong signal component is received by virtue of reflection from the water surface of Raritan Bay. Under normal conditions, the geometry of the path does not permit such a reflection.

Normally the dielectric constant of the atmosphere decreases with height above ground so that the ray path usually has a curvature in the same direction as the earth curvature. However, it is possible for the dielectric constant of the atmosphere to increase with height above ground (sub-standard conditions) so that the ray path has a curvature opposite that of the earth. This results in the condition illustrated in Fig. 3(d) where the limiting or tangent ray does not reach the receiver and only a weak signal is received by virtue of diffraction. Widely separated frequencies and vertically spaced antennas are affected alike as regards the average signal level but not the fine structure fading. This effect has been observed only on the Southard Hill-Crawford Hill path which has small clearance to begin with. It has been observed on several nights in late summer or early autumn after a radiation type ground fog has formed in the late evening and usually persists until the fog is dispelled by winds or by the morning sun.

There are, of course, times when the transmission conditions are considerably more complicated than those described above. Some of these apparently are due to a combination of the situations illustrated in Fig. 3 while others may be the result of an atmospheric focussing or trapping phenomenon. In addition to the various phenomena just described, which, fortunately, occur rather infrequently, there are numerous occasions when the signal varies plus and minus a few decibels relative to the free space level. It has not been possible actually to demonstrate the mechanism responsible but it seems most likely that these smaller variations are due to non-linear dielectric constant gradients which give the atmosphere the properties of a convergent or divergent lens.

An important result of the observations made to date is the conviction that the severe fades, signal excursions to levels 30 decibels or more below the free space field, were all caused by wave interference. It appears that, as the average signal level is depressed by any mechanism, it becomes more and more vulnerable to the effects of extra signal components of small amplitude that often may be present but go unnoticed when the signal is near normal levels. Thus, while the average signal level during the conditions illustrated in Figs. 3(b) and 3(d) may be no more than 15 to 20 decibels below the normal daytime level, there

is usually superimposed a fine structure fading in which short duration fades to levels as much as 45 decibels below free space have been observed. For this reason it is desirable to avoid paths having small clearance over intervening terrain and also paths which have a permanent ground reflection of sufficient magnitude to depress the signal to critical levels when, due to variable atmospheric refraction, the direct and reflected components are in phase opposition.

The following sections describe the angle-of-arrival and frequency-sweep experiments on which much of the preceding discussion was based.

ANGLE-OF-ARRIVAL OBSERVATIONS

A photograph of the Crawford Hill receiving site is shown in Fig. 4. The building housing the receiving equipment and the associated antennas are mounted on a framework which can be rotated on a concrete track, permitting investigation of the transmission characteristics of either path. The parabolic antennas on the tower are used for continuous recording of 4195 megacycle signals. The long object at the left of the

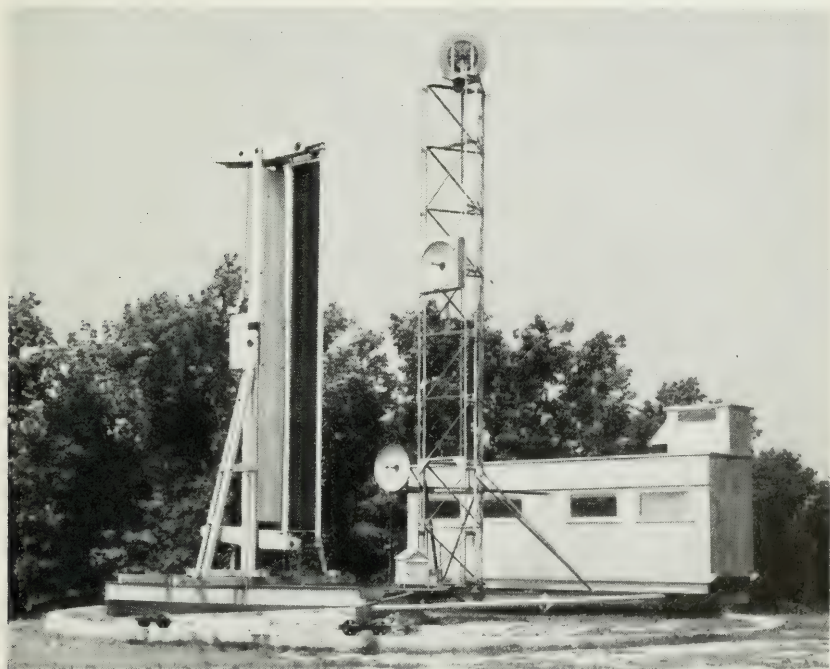


Fig. 4—The Crawford Hill receiving site.

picture is the metal-lens antenna used for making the angle-of-arrival observations. Its half-power beamwidth is 0.12 degree at the operating frequency of 24,000 mc. The focal length of the lens is 48 feet and its feed is located in the little cupola on top of the building. The feed is held fixed, while the lens is moved vertically by a motor-driven mechanism; thus the antenna beam also moves vertically. The antenna scans a total angle of two degrees in ten seconds. It is fed by a 24,000-mc radar set which is gated to receive only the pulses reflected from a corner reflector located at the distant terminal of the transmission path. The spot on the radar cathode ray tube moves vertically in synchronism with the scanning antenna, and the horizontal deflection is proportional to the amplitude of the pulse received from the corner reflector. The display thus shows amplitude of the various incoming signal components as a function of their angles of arrival.

The antenna installation on Southard Hill is shown in Fig. 5. At the left is the transmitting paraboloid for the 4195-mc continuous wave transmitter, the radar corner reflector is in the center, and on the right is the horn-reflector antenna used in the frequency-sweep experiments described below. Similar equipment is located at the Murray Hill terminus. The corner reflector is 5.5 feet on a side, and at 24,000-mc has sufficient gain to override reflections from other nearby objects, and thus becomes easily identifiable on the radar screen.

The radar oscilloscope for typical propagation conditions is shown in Fig. 6. These pictures were obtained by leaving the camera shutter open during the ten-second interval required for the antenna beam to scan through the angular range of 2° . All of these representative photographs were taken on the Murray Hill-Crawford Hill path although similar results were obtained on the Southard Hill-Crawford Hill path with the exception of Fig. 6(f). The normal daytime transmission is shown in Fig. 6(a) to consist of a single path arriving at an angle of -0.2° with respect to a fixed reference angle. The horizontal lines represent intervals of 0.1° , so that changes of 0.05° can be estimated. The other pictures in Fig. 6 were all taken during fading conditions.

Figs. 6(b) and 6(c) are good examples of the multiple-path condition shown in Fig. 3(a) in which the individual components are almost equal in amplitude and well separated in angle. In Fig. 6(b) there are two components arriving at angles of 0.1° and 0.6° above the normal line-of-sight while in 6(c) there are three components with angles of 0.05° , 0.35° and 0.7° above the normal angle. The position and amplitude of the signal components may change radically in a matter of minutes, and often there is no component that can be identified as the "normal" one.



Fig. 5—The Southard Hill transmitting site.

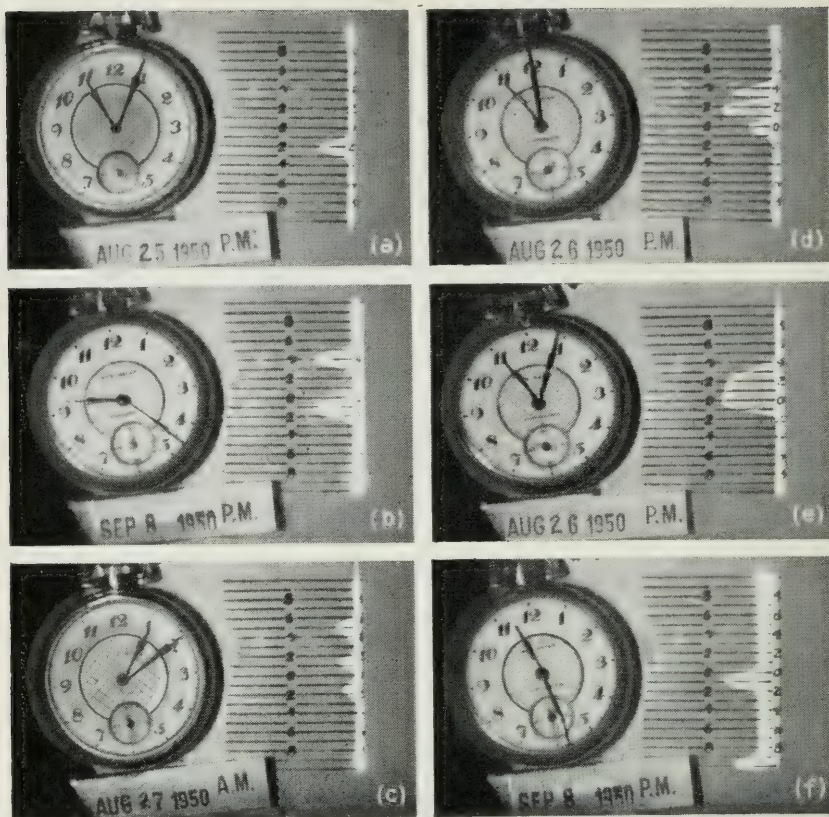


Fig. 6—Representative photographs of angle-of-arrival observations on the Murray Hill-Crawford Hill path. (a) Normal day. (b) Two elevated paths. Sept. 8, 1950; 9:23 p.m. (c) Three elevated paths. Aug. 27, 1950; 1:11 a.m. (d) Multiple paths. August 26, 1950; 11:00 p.m. (e) Wide angle "fill-in". Aug. 26, 1950; 11:04 p.m. (f) Abnormal water reflection. Sept. 8, 1950; 11:28 p.m.

During these multiple-path conditions, the recordings of the 4195-mc transmission generally show the broad maxima and sharp minima characteristic of wave interference.

Figure 6(d) shows a case in which the various paths are not completely separated while Fig. 6(e) (taken four minutes later) shows that energy is being received almost without variation over a vertical angle of 0.4° . This may represent a number of ray paths which would be separable by a narrower-beam antenna, or it may indicate a focussing or trapping phenomenon. Often when the type of transmission illustrated by 6(e) is present, the recorded 4195-mc signal may be as much as 12 to 15 decibels above the free space levels.

Fig. 6(f) illustrates the case of abnormal reflection from the water of Raritan Bay on the Murray Hill path as indicated in Fig. 3(c). Here the "normal" signal component is arriving at 0.1° above the line-of-sight while another component, almost equal in amplitude, is arriving at the very bottom of the scan, about 0.8° below the line-of-sight. It is quite probable that there have been times when this component was present but was outside the range of the scanning antenna.

The mechanisms discussed in connection with Fig. 3(b) and 3(d) cannot be demonstrated by photographs such as those just presented although the angle-of-arrival radar was instrumental in furnishing the clues to the phenomena. Due to the two-way attenuation of the radar-corner reflector technique, the signal at these times rapidly falls below the noise level of the receiver. For the same reason, it is not possible to detect the extra signal components of small amplitude which were postulated to account for the very deep fades sometimes observed under these transmission conditions.

FREQUENCY-SWEEP OBSERVATIONS

Since most of the fading is due to interference between waves which travel over different paths of, presumably, different lengths it was realized that the fading was likely to be frequency selective. Just how selective would depend on the relative lengths of the individual transmission paths. The usual methods for determining path length differences are to use short pulses, or to sweep the frequency. Since it was likely that the path-length differences would be measured in feet rather than yards, very short pulses or a wide frequency-sweep were required. An oscillator³ was available whose frequency could be swept over the licensed band of 500 mc between 3700 mc and 4200 mc. The frequency-sweep experiment was set up on the Murray Hill-Crawford Hill path for the summer of 1949. The following summer, the milli-microsecond pulse transmission tests described in the companion paper were conducted over the same path. As might be expected, simultaneous observations showed good agreement between the two methods.

The frequency of the transmitter, located at Murray Hill, is swept over a 450-mc band centered at 3950 mc at a 60-cycle rate. At the receiver, a similar oscillator is used for the beating oscillator except that its frequency is swept linearly through the same frequency band in one

³ This oscillator was developed by M. E. Hines and is described in his paper published in the *Bell System Technical Journal*, Vol. 29, Oct. 1950. It uses a 416A close-spaced triode in a wave-guide cavity. The frequency is changed by means of a plunger which is capacity-coupled to the plate of the tube and which is actuated by a modified loud speaker unit.

second. Since the intermediate frequency amplifier of the receiver is only 350 kc wide, (centered at 600 kc) narrow pulses are generated each time the frequency of the transmitter crosses the frequency to which the receiver is tuned. These intermediate frequency pulses are displayed vertically on a cathode ray tube. The horizontal trace is synchronized with the one-second sweep rate of the beating oscillator.

The normal daytime frequency-sweep pattern is shown in Fig. 7(a). The vertical scale is linear in amplitude and the horizontal scale is almost linear in frequency, with frequency decreasing from left to right. Visible at the extreme left is the signal used for continuous recording. Since there is only one transmission path involved, the amplitude of the received signal is nearly constant over the 450-mc band. If another signal were present which had travelled over a path of different length, the two signals would add when the frequency is such that the path length difference is an even multiple of half-wavelengths and subtract when the path length difference is an odd multiple of half-wavelengths. Simple calculation shows that if the path length difference is one foot, the frequencies at which the signals add and subtract are separated about 500 mc. Thus the limit of resolution for the frequency-sweep experiment is a little more than one foot.

Photographs taken on a night when the angle-of-arrival radar indicated two almost equal components separated about 0.4 degrees in angle are shown in Fig. 7(b). The time interval between the two pictures is 30 seconds, during which the minimum had shifted about 150 mc. The pictures can be interpreted as simple two-path transmission with an indicated path difference of about two feet and an amplitude ratio of 0.7 to 1. On this night the minimum shifted back and forth across the frequency band—sometimes slowly and sometimes rapidly. At times the position of the minimum might remain fixed but its depth would change.

Photographs taken on a night when there were abnormal reflections from the water of Raritan Bay are shown in Fig. 7(c). There are evidently two main components with path difference of about six feet, with a small third component causing the slight decrease in amplitude of the peaks from left to right. These pictures were taken 9 minutes apart, but this type of pattern was observed over a period of about three hours on this night.

Usually the frequency sweep patterns are considerably more complicated than the ones shown so far. Fig. 7(d) shows two photographs which indicate that at least three signal components and perhaps more were present. The time interval between the two pictures was about 30 seconds.

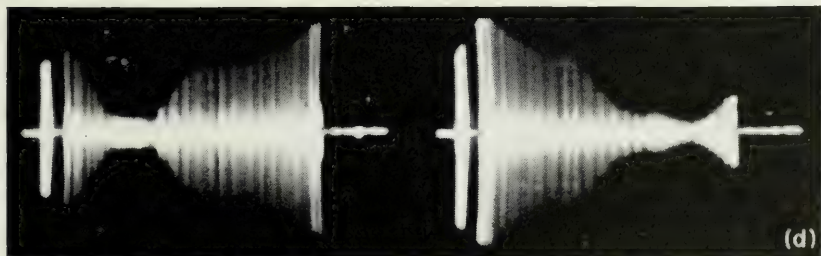
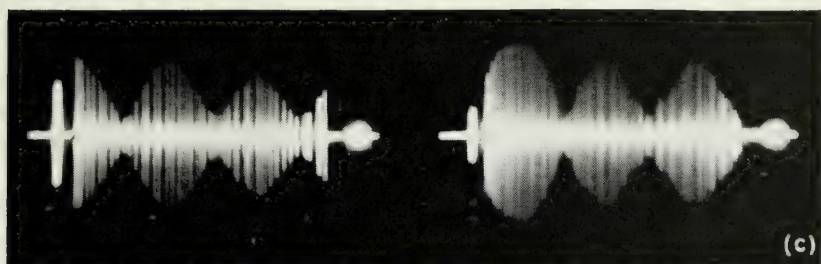
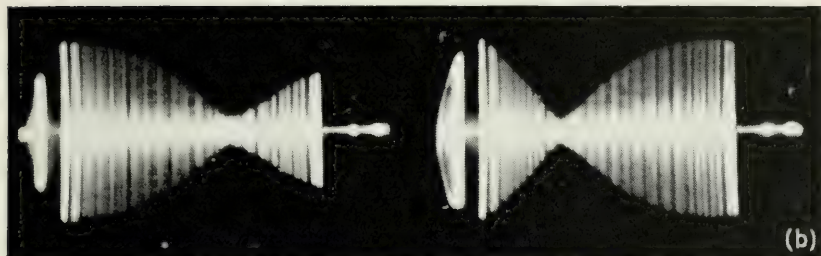
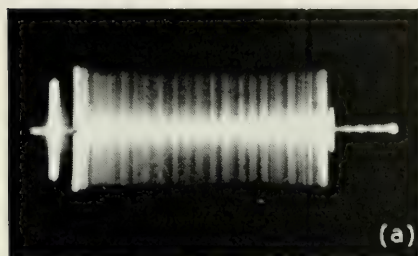


Fig. 7—Representative frequency-sweep patterns observed on the Murray Hill-Crawford Hill path. (Summer 1949.) (a) Normal day. (b) Two components with a path difference of two feet. (c) Two main components with a path difference of about six feet plus a small third component. (d) Multiple component pattern.

SYNTHESIS OF FREQUENCY-SWEEP PATTERNS

To aid in the interpretation of the complicated frequency sweep patterns, a computer of the analogue type was built. This apparatus combines four signal components, three of which are variable in delay and amplitude, and presents the result on a cathode ray tube in the same form as the actual frequency sweep patterns. Thus a particular pattern can be synthesized on the computer and the number of components, together with their path differences and relative amplitudes, read directly from the computer dials. This is accomplished by generating four 600-kc signals, three of which are phase modulated at 60 cycles per second. The total phase deviation and relative signal amplitude are variable. The four signals are then summed and displayed in vertical deflection on a cathode ray tube having a 60-cycle horizontal sweep.

The synthesis of the patterns of Figures 7(b) and 7(c) are shown in Fig. 8. The upper synthesized pattern is simply a combination of two components with relative amplitudes of 0.7 and 1 and a path difference of two feet. The lower pattern consists of the reference component with unity amplitude, a second component with an amplitude of 0.5 and a path difference of 5.7 feet, and a third component with an amplitude of 0.2 and path difference of 0.8 feet. The similarity between the actual and synthesized patterns is obvious.

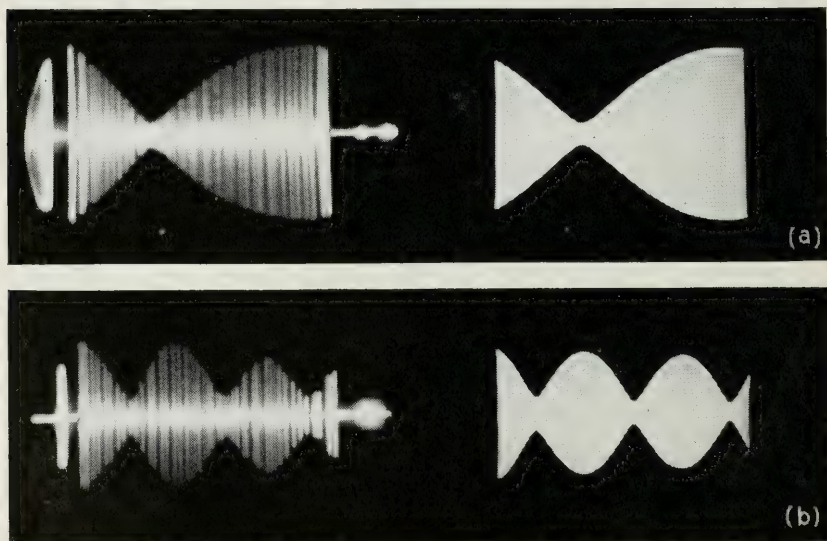


Fig. 8—Synthesis of the frequency-sweep patterns of Figs. 7(b) and 7(c).

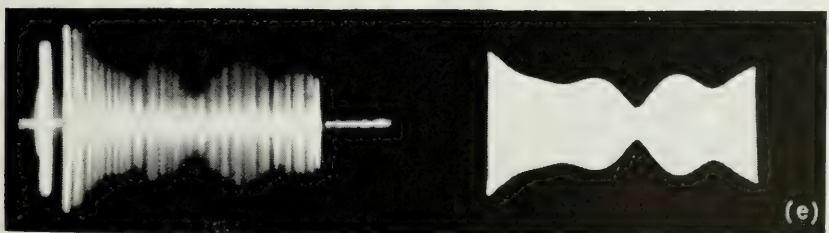
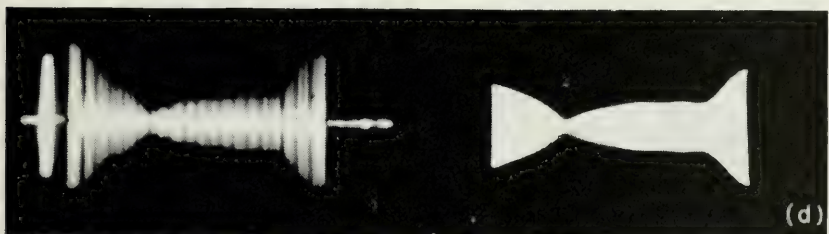
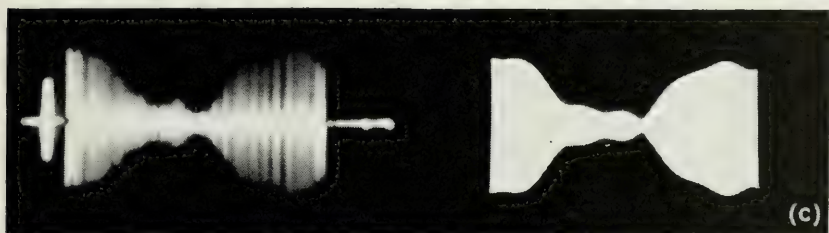
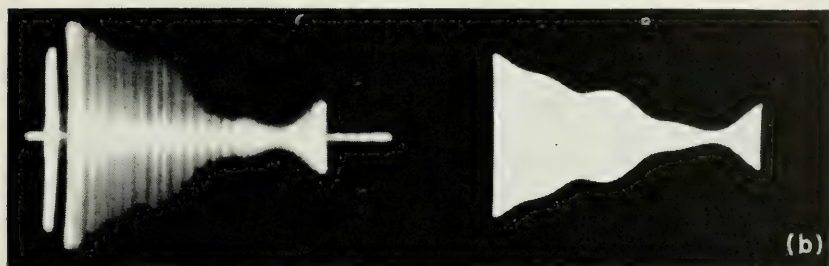
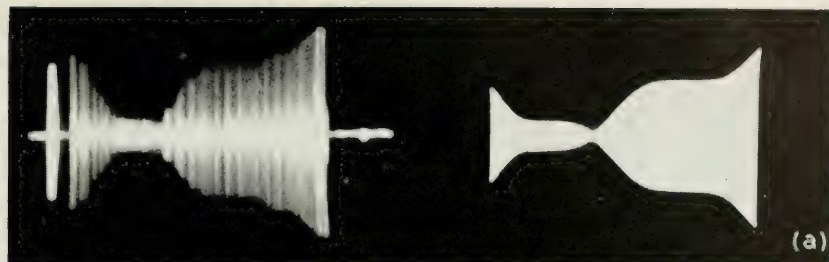


Fig. 9—Synthesis of complicated frequency-sweep patterns using four components. See Table I for values of the relative amplitudes and path differences.

Examples of attempts to synthesize some more complicated frequency sweep patterns, taken on the night of August 2, 1949, are shown in Fig. 9. Four components were required in each case, with the path differences and delays being summarized in Table I. Although the pictures all appear very different, in general major changes were required only in component No. 2 to go from one pattern from the next, as Table I shows. All the remaining components had to be very carefully trimmed in both amplitude and delay to get good synthesis (especially in the case of Fig. 9(d), but these changes were relatively small.

CONCLUDING REMARKS

The special experiments just described have led to the conclusion, expressed earlier, that the severe fading observed on the two test paths is the result of multiple-path transmission in which several components may be involved. These components may arrive at the receiver at various angles up to three quarters of a degree above the normal daytime angle-of-arrival and, in the case of abnormal water reflection on the Murray Hill path, as much as 0.8 degree below the normal angle. The path differences among these components may vary from a fraction of a foot to about ten feet. The long-delay components are usually small in amplitude.

In all cases where observations were made during periods of exceptionally high signal levels, say 10 to 15 decibels above free space level, the frequency-sweep patterns were substantially flat, suggesting a focusing or trapping phenomenon. The frequency-sweep patterns were also flat on those nights when the signal excursions were only a few decibels above and below the normal daytime level. However, the severe fades

TABLE I

		Fig. 9 (a)	Fig. 9(b)	Fig. 9(c)	Fig. 9(d)	Fig. 9(e)
Component No. 1 (Reference)	Amplitude	1	1	1	1	1
	Path diff. (ft.)	0	0	0	0	0
Component No. 2	Amplitude	0.9	1.2	0.7	1.1	0.4
	Path diff. (ft.)	1.1	0.5	1.7	0.5	1.1
Component No. 3	Amplitude	0.2	0.1	0.2	0.2	0.2
	Path diff. (ft.)	5.2	5.7	5.6	5.7	5.7
Component No. 4	Amplitude	0.05	0.1	0.15	0.1	0.1
	Path diff. (ft.)	9.2	9.2	11.0	9.3	8.7
Time		12:08½ AM	12:09 AM	12:18 AM	12:24 AM	12:25 AM

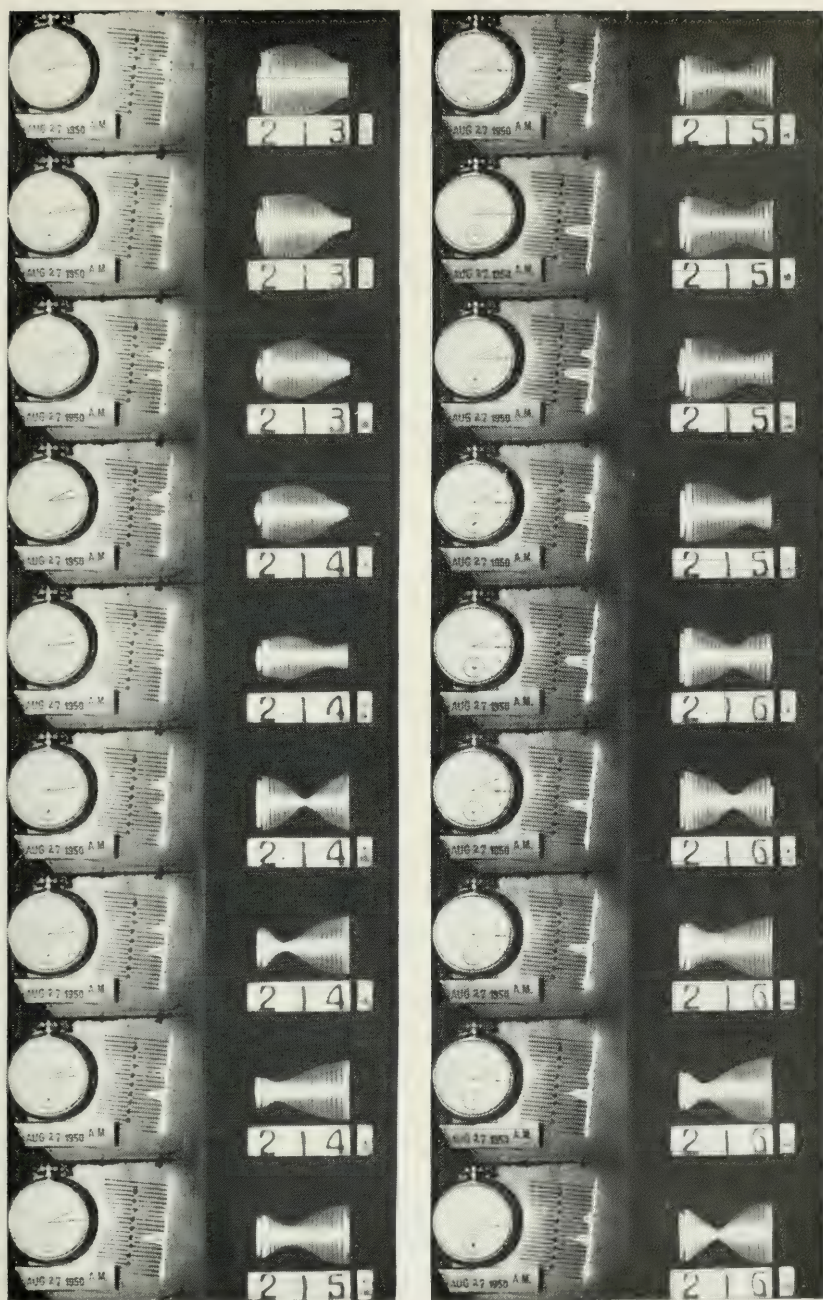
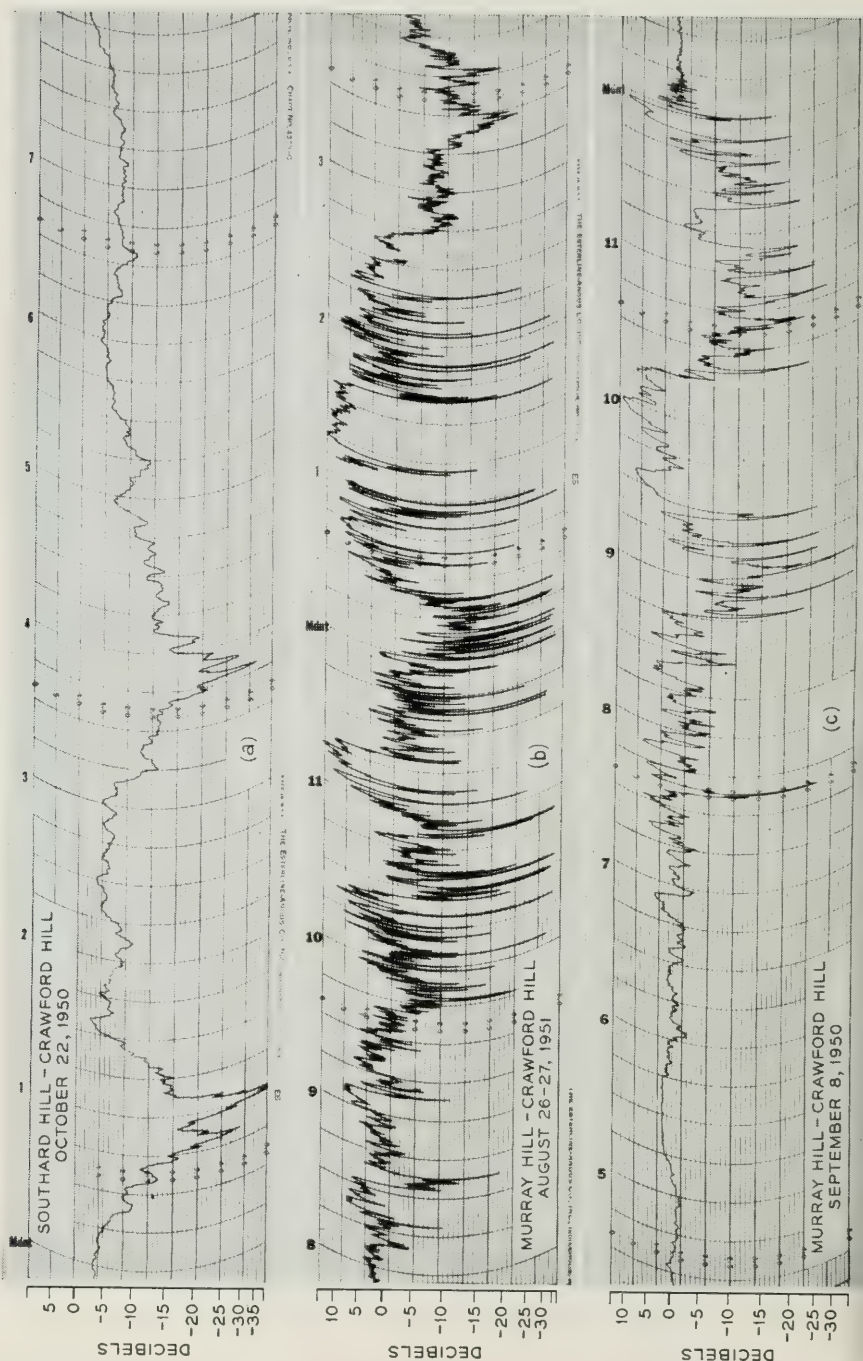


Fig. 10—A sequence of angle-of-arrival and frequency-sweep patterns taken at ten-second intervals.



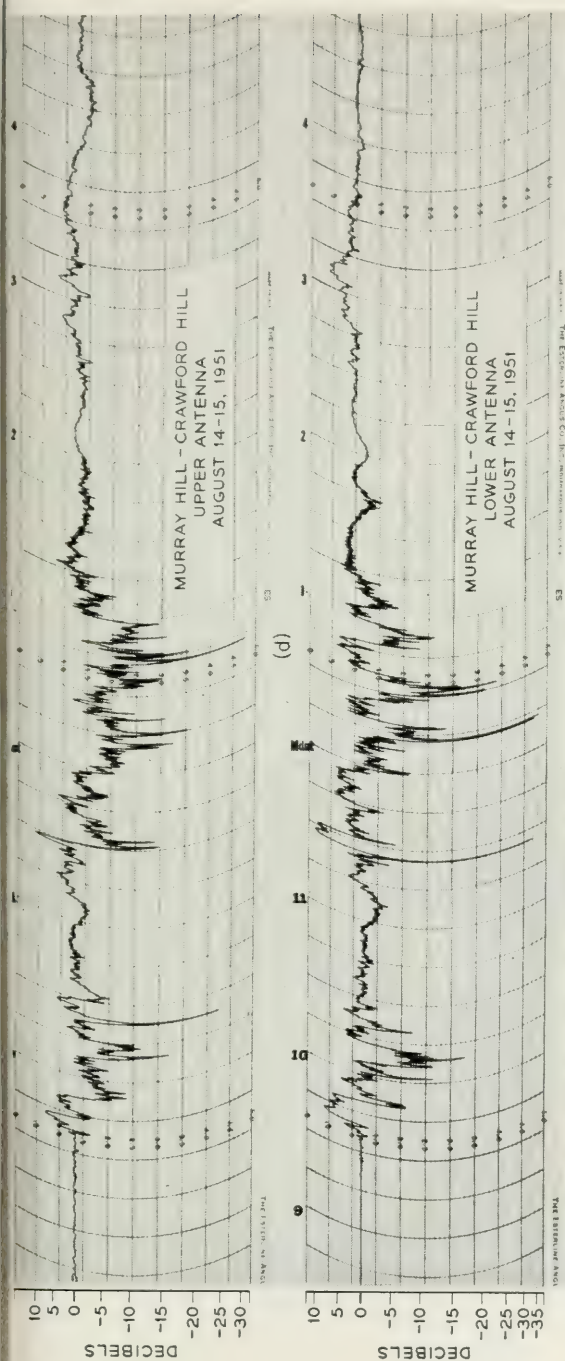
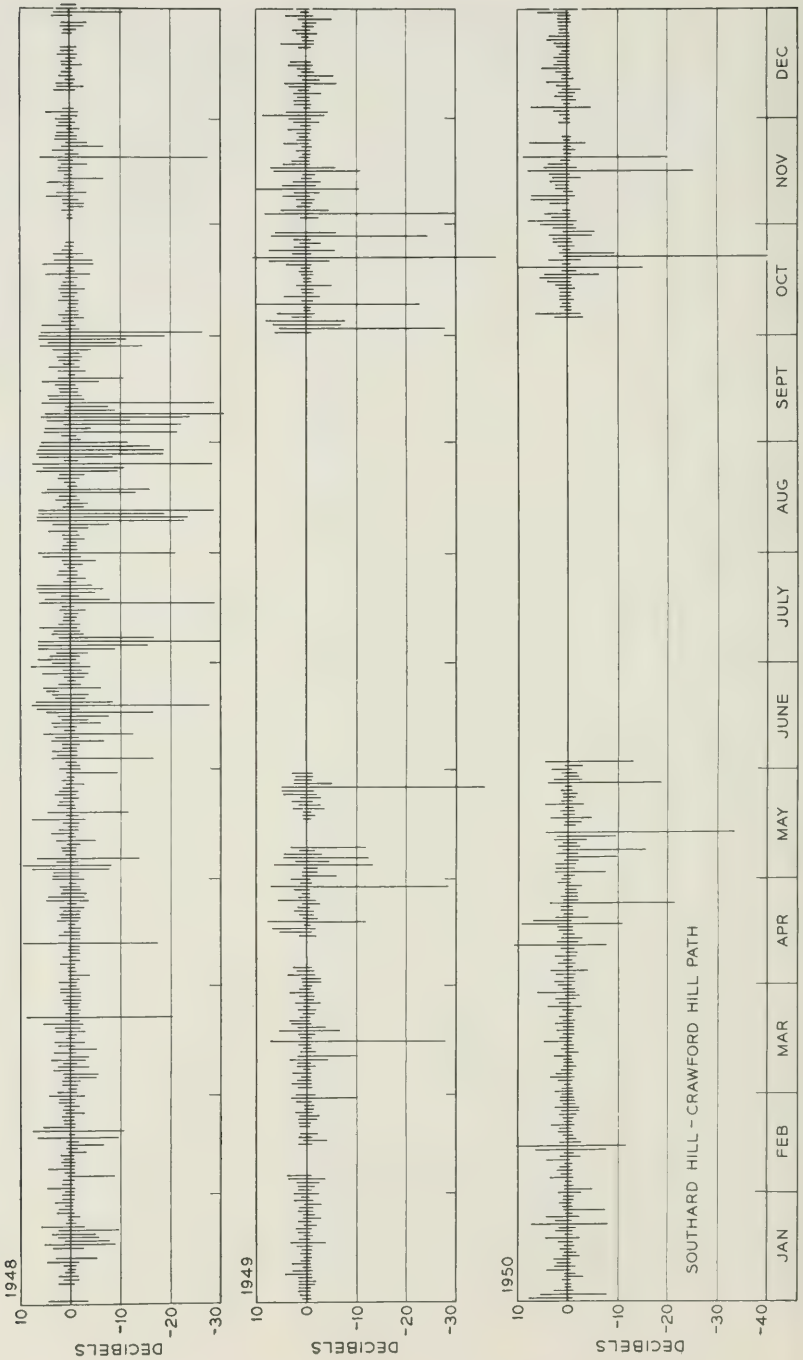


Fig 11.—Typical recordings of 4195-mc transmission during severe fading conditions. The zero decibel line represents the normal daytime (free space) level. (a) Substandard conditions on the Southard Hill-Crawford Hill path. Night of Oct. 21-22, 1950. (b) Multiple path transmission on the Murray Hill-Crawford Hill path. Night of Aug. 26-27, 1950. (c) Multiple path transmission on a night when abnormal water reflections were present on the Murray Hill-Crawford Hill path. Night of Sept. 8-9, 1950. (d) Simultaneous recording of the outputs of two similar antennas; vertical spacing of 30 feet. Night of Aug. 14-15, 1951.



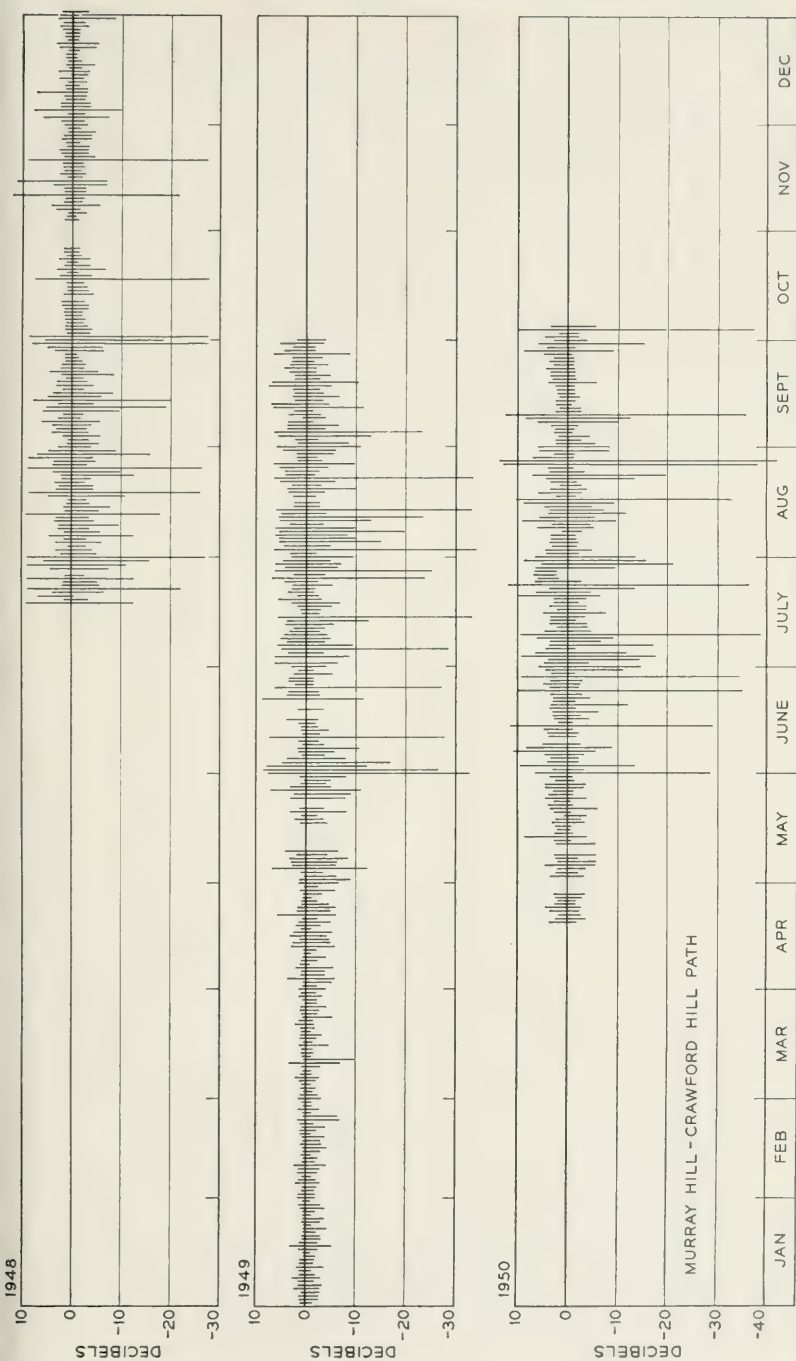


Fig. 12—Chart showing the daily fading range observed on both transmission paths for the years 1948, 1949, and 1950. The vertical lines terminate on points indicating the highest and lowest signals observed during the twenty-four hour period of noon to noon.

of short duration sometimes observed when the average signal level was depressed by the mechanisms of Figs. 3(b) or 3(d) were found to be frequency selective.

Some of the studies described in this paper were made with vertically polarized waves and some with horizontally polarized waves; at times, 45° polarization was used. In so far as it was possible to determine, the propagation characteristics of both paths were independent of the polarization used.

No meteorological soundings were made in connection with this work. Considering the rapid changes usually observed with the angle-of-arrival and frequency sweep apparatus, it is doubtful that meteorological measurements made in the usual manner would show much correlation with the radio observations except, perhaps, in a general way. The sequence of pictures in Fig. 10 is included to show how the angle-of-arrival and

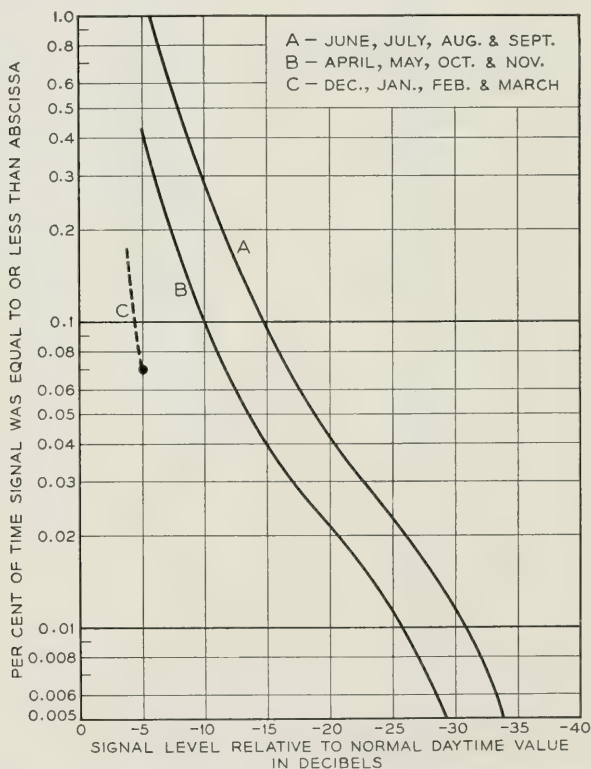


Fig. 13—Time distribution curves of the signal levels observed on the Murray Hill-Crawford Hill path. Data of 1947, 1948, 1949 and 1950.

frequency-sweep patterns change with time. These pictures were taken at 10-second intervals. On this occasion there was good correlation between the angle-of-arrival and frequency-sweep data. Such was not always the case, however, and considering the wide difference in operating frequencies, 24,000 mc and 4000 mc, instantaneous correlation should not necessarily be expected.

Although all the studies described in this paper were made on the two local paths, the results are compatible with propagation measurements made by another group in the Laboratories during a survey for the transcontinental radio relay system.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the contributions of W. M. Sharpless who, for some time, was associated with this work; also to acknowledge

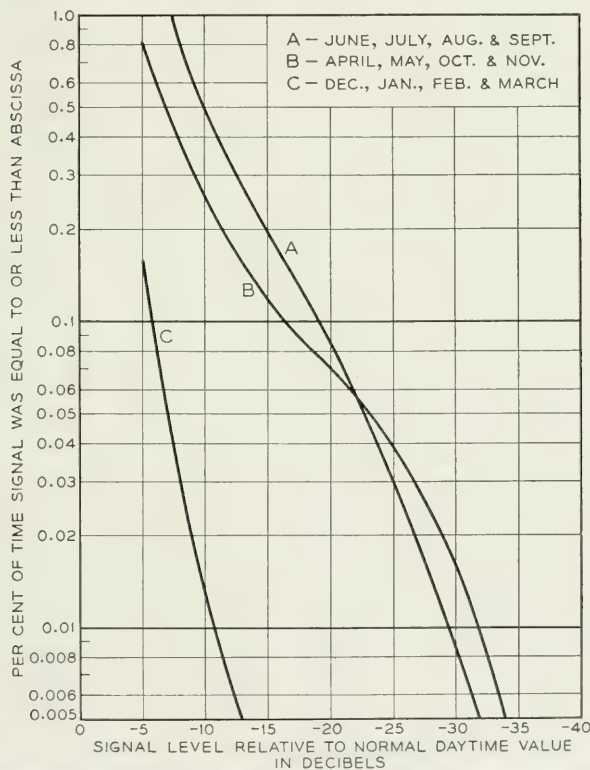


Fig. 14—Time distribution curves of the signal levels observed on the Southard Hill-Crawford Hill path. Data of 1947, 1948, 1949 and 1950.

the full time assistance of R. A. Desmond and the part time assistance of L. R. Lowry and S. E. Reed. All the work was done under the guidance of Dr. H. T. Friis.

APPENDIX

This appendix is included to illustrate some of the characteristics of the propagation as shown by the recordings of 4195-mc signal levels. Fig. 11 is a reproduction of some typical records obtained during severe fading periods. Fig. 11(a) is an example of transmission during the substandard conditions illustrated by the ray diagram of Fig. 3(d). Fig. 11(b) is typical of multiple-path type fading in which the signal components arrive from elevated angles as shown in Fig. 3(a), while Fig. 11(c) was recorded on a night when, for a time, there were abnormal reflections from the water of Raritan Bay on the Murray Hill path, see Figs. 3(e), 6(f) and 7(c). The records of Fig. 11(d) show how the outputs of two similar antennas, spaced vertically about 30 feet, differ in regard to the deep fades of short duration.

The chart of Fig. 12 shows how the fading varies with the time of year. On this chart, the vertical lines represent the extremes in signal level observed during the twenty four hour period from noon to noon. The large signal variations are concentrated mainly in the summer months.

The time distribution of the signal levels recorded on the Murray Hill-Crawford Hill path are shown in Fig. 13. Each of the curves is for a four-month period: the period of least fading, December, January, February and March; the period of most fading, June, July, August and September; and the in-between period consisting of April, May, October and November. Data obtained in the years 1947, 1948, 1949 and 1950 are included. Fig. 14 shows similar data for the Southard Hill-Crawford Hill path. The hump in the time distribution curve for the months of April, May, October and November is due to substandard conditions, illustrated by the ray diagram of Fig. 3(d) and the typical record of Fig. 11(a), which affected transmission on this path during several nights in October, particularly in the years 1947 and 1950. When it occurred, this type of transmission usually persisted for a period of several hours.

Propagation Studies at Microwave Frequencies by Means of Very Short Pulses

BY O. E. DE LANGE

(Manuscript received March 27, 1951)

Microwave pulses with a duration of about 0.003 microseconds were transmitted over a 22-mile path from Murray Hill, N. J., to Holmdel, N. J., in order to determine the effects of the transmission medium upon such pulses. During "fading" periods multi-path transmission effects with path differences as great as 7 feet were observed, as well as some other effects. A microwave frequency of 4000 megacycles was employed.

INTRODUCTION

This experiment was set up with two main purposes in view: First, as a means of studying microwave propagation, especially with regard to multi-path transmission effects and second, to determine the effect of a transmission path upon the shapes of very short pulses, particularly to learn what restrictions might be imposed upon minimum pulse length or spacing between pulses by distortions produced in the transmission medium.

In regard to multi-path transmission the pulse method seems to be the most straightforward way of studying such effects. For example, if there is transmission by more than one path, and if the pulses are sufficiently short in comparison to the path length differences involved, then there will be received a separate pulse for each path. Under these conditions the number of paths involved, path length differences and other information become directly evident. If pulse duration is too great with respect to the path differences involved, the pulses received via the various paths will overlap in time and the resultant multi-path effect will be pulse distortion rather than reception of individual pulses. This situation is much more difficult to analyze.

TRANSMISSION PATH

The transmission path is the same as that used by A. B. Crawford for microwave propagation studies by means of the frequency sweep method,

i.e., the path from Murray Hill, N. J., to Crawford's Hill (near Holmdel), N. J.¹ The path length is approximately 22 miles, and is partly over water and partly over rough land terrain. The frequency sweep studies had indicated that the path differences involved in multi-path fading were of the order of one or two to about seven feet. In terms of delay times this means differences of about 1 to 7 millimicroseconds. In order to resolve the paths when the path differences were only one or two feet, we should have liked to have pulses of about 1 millimicrosecond duration. Because of the difficulties involved in generating, amplifying

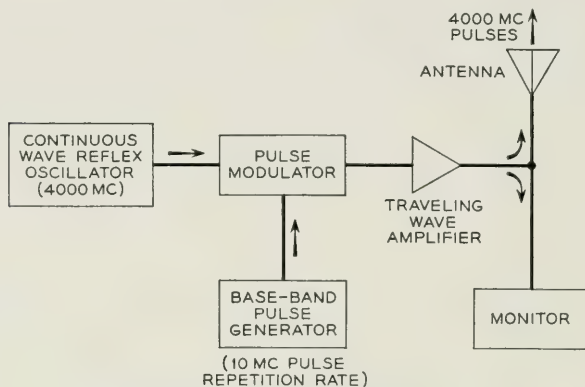


Fig. 1—Transmitting equipment.

and detecting such short pulses, we accepted pulses which, when displayed on our final indicating equipment, had a length of 3 millimicroseconds at half amplitude. (About 6 millimicroseconds at the base.) In free space this pulse would be just about 6 feet long at the base.

TRANSMITTING EQUIPMENT

The transmitter was mounted on top of a 100-foot tower at Murray Hill. As can be seen from Fig. 1, it consisted of a c-w reflex oscillator operating at 4000 megacycles, a baseband pulse generator, a modulator, or gate, for modulating these pulses on the microwave carrier, a single stage traveling-wave amplifier and finally a horn antenna. Approximately one watt of power was obtained from the transmitter at the peaks of the pulses. The antenna area was 25 square feet and its gain 32 db above that of a dipole. A pulse repetition frequency of 10 mc was employed.

¹ A. B. Crawford and W. C. Jakes, Jr., "Selective Fading of Microwaves," *Bell System Tech. J.*, **31**, Jan. 1952, pp. 68-90.

RECEIVING EQUIPMENT

The receiving antenna, a large horn, was mounted between two poles guyed for support. It had an aperture of about 90 square feet and a gain of approximately 38 db over a dipole. The receiver circuit is shown in Fig. 2. About 60 db of gain at 4000 mc was provided by either two or three stages of traveling-wave tube amplifier depending upon the gain of the particular tubes used. It was necessary to provide very good shielding and also careful filtering of all power leads to eliminate the tendency for this amplifier to sing. The amplifier fed two crystal detectors through a hybrid tee junction. Each detector employed a silicon crystal of the IN23B type.

Two indicator circuits are shown in Fig. 2. These circuits are very similar except that one employed a vertical amplifier coupled to a Dumont 5XP2 CRO tube, whereas in the second the baseband output of the crystal was fed directly onto the deflection system of a traveling-wave type of CRO tube. The latter CRO tube, which has been described by J. R. Pierce in the November, 1949, issue of *Electronics*, has a very high deflection sensitivity and is used with a microscope to enlarge its trace; hence, no amplification was required between it and its driving crystal. The deflection system of this tube has a bandwidth of 500 to 1000 mc. The micro-oscilloscope was provided primarily for photographing pulses by means of a 35-mm camera attached to the microscope. (Exposure time was 5 to 15 seconds. The time recorded for each picture corresponds

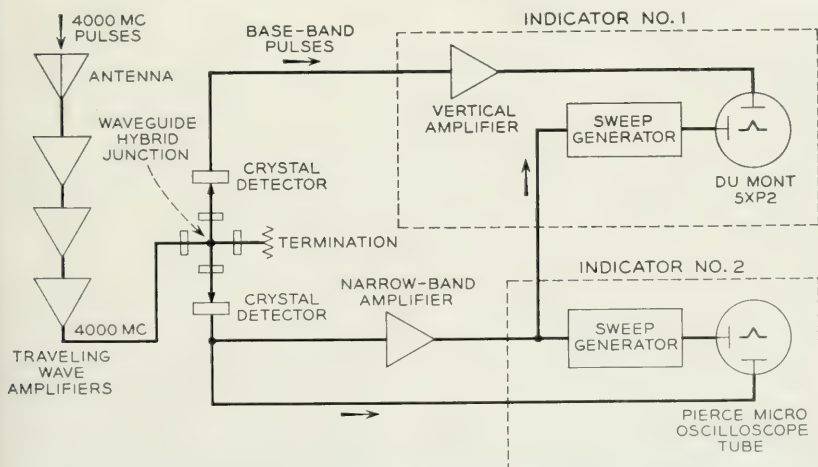


Fig. 2—Receiving equipment.

to that at the end of exposure.) A second microscope made it possible to view and to photograph the screen of the tube simultaneously. The general procedure was to observe continuously during periods of disturbed transmission, taking pictures at regular intervals of 5 to 10 minutes. When conditions were seen to be changing rapidly, pictures were taken much more frequently. The large oscilloscope with its vertical amplifier had a bandwidth of about 150 mc and hence caused some deterioration of the pulse. It, however, was less tiring than the small scope, especially for long periods of observation and was watched to follow the general trend of events. It was capable of resolving the pulses resulting from two-path transmission when the path differences were large.

The sweep circuits for the two indicator oscilloscopes were practically identical. The horizontal sweep voltage for each consisted of the linear portion of a sine wave which was generated by a c-w oscillator operating at one third of the pulse repetition frequency of 10 mc. Each oscillator was synchronized with the incoming pulses by means of a 10-mc voltage derived by amplifying the pulse energy through a narrow band amplifier. This circuit provided very satisfactory synchronization even during the times when signal amplitude was so low as to produce a very poor signal-to-noise ratio. Timing markers were provided on each roll of film by periodically photographing a series of pulses spaced by an interval of 9 millimicroseconds.

RESULTS OF THE EXPERIMENT

The picture at the left of Fig. 3 shows the transmitted pulse. The right-hand picture shows the received pulse under what were considered to be normal transmission conditions. It is seen that, except for the addition of noise and widening of the pulse due to passage through the amplifiers and other equipment, the pulse shape is unaffected. The time calibration on this and the following photographs are in millimicroseconds, each mark representing one millimicrosecond ($0.001 \mu\text{s}$).

During the summer of 1950, when this experiment was in progress,

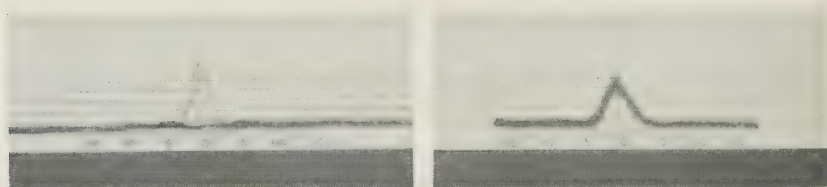


Fig.3—(Left) transmitted pulse (right) received pulse—normal transmission.

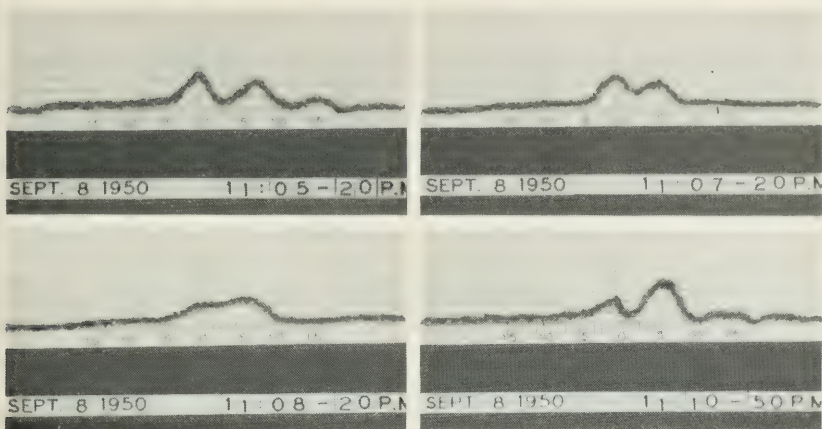


Fig. 4—Received pulses—disturbed transmission.

there was comparatively little fading over the path in question at microwave frequencies. There were, however, a few nights of considerable activity and some interesting results were obtained.

The series of pictures on Fig. 4 show one good example of multi-path transmission where the path length difference was great enough to produce complete resolution of pulses. At 11:05-20 there are two pulses, each 7 millimicroseconds wide at the base and with their peaks just 7 millimicroseconds apart; in other words, the path difference was just sufficient to produce two pulses with no overlap. The pulse at the left is presumably coming by the main path and that at the right from some second path resulting from bending of the rays caused by atmospheric effects. At 11:07-20 the second path appears to have shortened, resulting in a path difference of only about 5 millimicroseconds. This may actually have been due to a change of length of the second path or it may have been due to distortion of the second pulse by energy coming by way of a third path. The pictures taken at 11:08-20 and 11:10-50 show evidence of transmission by a third path. In the first of these, for example, the width of the disturbance at the base line indicates the presence of the two original pulses spaced 7 millimicroseconds apart but the midpoint of the two no longer falls to the base line as was the case in the first picture. This could be accounted for by the presence of a third pulse coming over a path whose length was somewhere between that of the other two. Conditions obtaining at 11:10-50 could also be accounted for by the presence of pulses from three paths, that is, energy coming by way of a third path might cancel part of one pulse and at the same time add to the other. This could account for the fact that the spurious pulse is larger than the

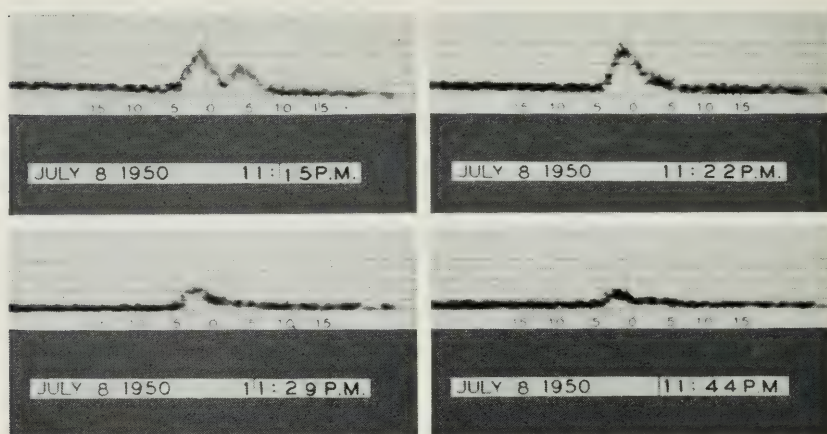


Fig. 5—Received pulses—disturbed transmission.

normal one. It is also possible that more than three paths were involved. On a number of other occasions the pulse coming by way of the second path appeared to be of greater amplitude than the one coming by the main path. This same effect has been observed by Mr. Crawford and his colleagues on the angle of arrival equipment.

Information obtained from the above set of pictures shows that for a time-division multiplex system using the length of pulse used here (7 millimicroseconds at the base) and operating over this path, pulses would have to be spaced a minimum of about 14 millimicroseconds apart if it were desired to avoid distortion at all times. If very much shorter pulses were used the spacing might be reduced to 9 or 10 millimicroseconds. However, the 7-foot path difference indicated by these pictures is about the maximum ever observed and occurs rather infrequently so that if somewhat closer spacings were employed troubles would result only a small percentage of the time.

The next series of pictures, Fig. 5, taken July 8, show an example of a more common type of multi-path transmission. Here the path difference is apparently less than for the last series. At 11:15 there are two distinct pulses with an apparent path difference of about six feet (6 millimicroseconds) if judged from the spacing between the peaks of the pulses. However, from the length of the disturbance at the base line, which we consider a better criterion, the path difference was more nearly four feet. At 11:22 distortion of the trailing edge of the pulse was the only indication of a second path. For the pictures taken at 11:29 and 11:44 the path difference is sufficiently small that there is almost complete cancellation of pulses, only the leading portion of each pulse being present.

On the 11:44 picture there is just a trace of a second pulse. The next set of pictures (Fig. 6) were taken a little over an hour later on the same night and show about the same conditions, that is, pulse amplitude and shape change and other evidence of the presence of a second pulse delayed about 2 to 3 millimicroseconds.

On the night of October 2, fading, which was apparently due to transmission by way of two paths with little path difference, was observed. Some of the results are shown on Fig. 7. At 7:49 two distinct pulses are evident, there being 6 millimicroseconds between their peaks. One might conclude from this that there was a second path about 6 feet longer than the main path but the total length of the disturbance along the base line and the shapes of the pulses indicate that the actual path difference was about 2 to 3 feet.

Apparently we had here two pulses of r-f energy overlapping in time and involving a large number of frequencies. These pulses are capable of interfering with each other in a rather complicated manner, it being possible for some frequencies to add and others to cancel at the same time, depending upon their relative phases. Phase relationships of course depend upon frequency and path length differences. As a result pulses may be distorted and have their peaks shifted about by a considerable amount. We must, therefore, realize that the first picture of Fig. 7 does not really represent two distinct pulses as appears to be the case, but actually shows the resultant interference pattern of two overlapping pulses. Since a change of path difference of only about one and one-half inches is enough to produce a 180° change in relative phase at 4000 mc, it is not

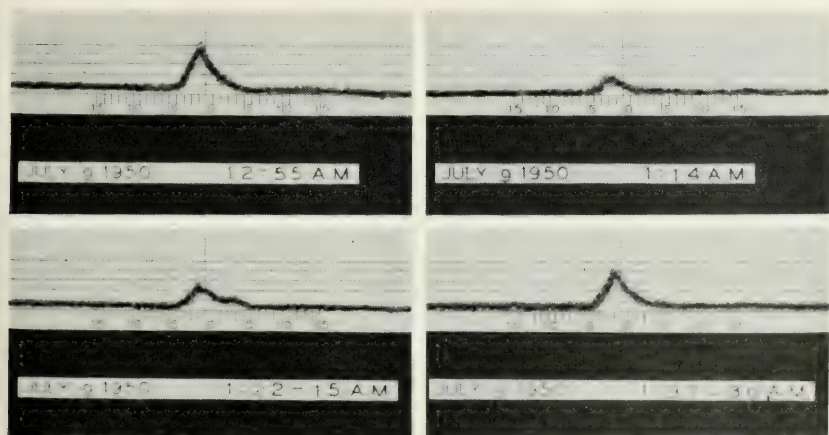


Fig. 6—Received pulses—disturbed transmission.

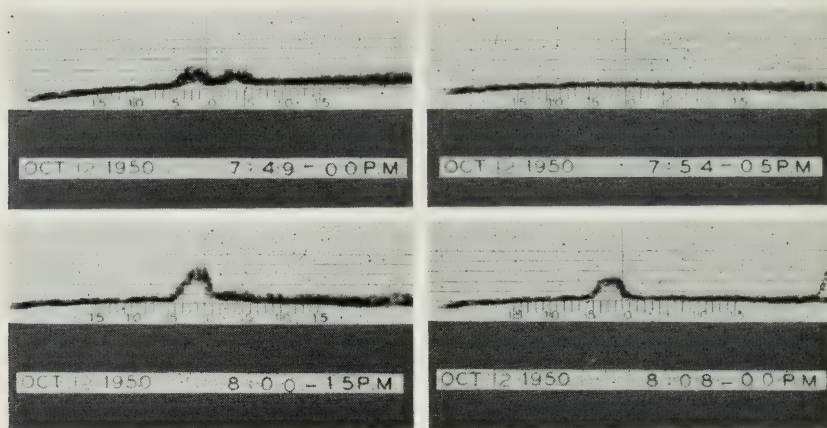


Fig. 7—Received pulses—disturbed transmission.

at all surprising that pulse shapes and amplitudes change very rapidly at times.

Looking again at the photograph, Fig. 7, we see that at 7:54-05 there was a complete fade as far as our system is concerned. To produce this degree of cancellation the path difference must have been very small though still sufficient to give a relative phase angle of 180° at the radio frequencies involved. At 8:00-15 and 8:08-00 pulse distortion is the most noticeable effect of the "fading," the pulses being considerably shorter than their normal value. Pictures, not shown here, taken between 7:54 and 8:08 show definite evidence of two-path transmission with a path difference of 2 to 3 feet; therefore the pulses of 8:00-15 and 8:08-00 are probably also the result of two-path effects.

The first two pictures of Fig. 8 show another form of pulse distortion observed on a number of occasions. Here the pulse is flattened out on top probably due to energy coming in over a second path differing in length by only one or two feet from the main path. Each time this type of pulse was observed a check was made to be sure that the flattening was not due to overload in our equipment. The pictures presented up to now have all shown comparatively slow changes of conditions. Very rapid changes were, however, quite common. In many cases pulse shape or amplitude changed considerably during the 5 to 15 second exposure time ordinarily used. The picture taken at 2:20-45 A.M. on August 27 is one example of such a rapid change, there being two definite sets of conditions shown on the one photograph. The remaining picture on Fig. 8 shows the pulses used for obtaining time calibration of the system. These pulses were spaced 9 millimicroseconds apart and by adjust-

ing sweep expansion so that succeeding pulses fell on proper parts of the scale and by keeping this expansion constant, it was possible to obtain a calibration.

TWO-PATH SIMULATOR

As an aid to interpreting the results obtained from the above experiment, particularly when the two pulses overlap and interfere, a circuit was set up in the laboratory to simulate two-path transmission. The equipment, as shown on Fig. 9, consisted of a wave guide hybrid junction with the r-f pulse energy being fed into the E plane arm. To each side arm was connected a variable attenuator in series with a few feet of wave guide fitted with a short circuiting plunger. Waves reflected from these two plungers recombine in the H plane arm where the detector is located. There are two separate paths through the hybrid as follows: (1) Input, side arm A, reflecting plunger A, side arm A to output. (2) Input, side arm B, reflecting plunger B, side arm B to output. By adjusting the attenuator in either branch the amplitude of the signal transmitted by way of that branch could be adjusted. In the same way by adjusting the position of the reflecting plunger in either branch the distance traveled by a signal in traversing that branch could be varied.

If the path lengths were made the same and the amplitudes adjusted to be equal there would be perfect cancellation due to a phase turn-over in the hybrid junction and hence no output from the detector. If one plunger were now left fixed in the above position and the other moved by a quarter wavelength (to produce a total shift of half wavelength or 180°)

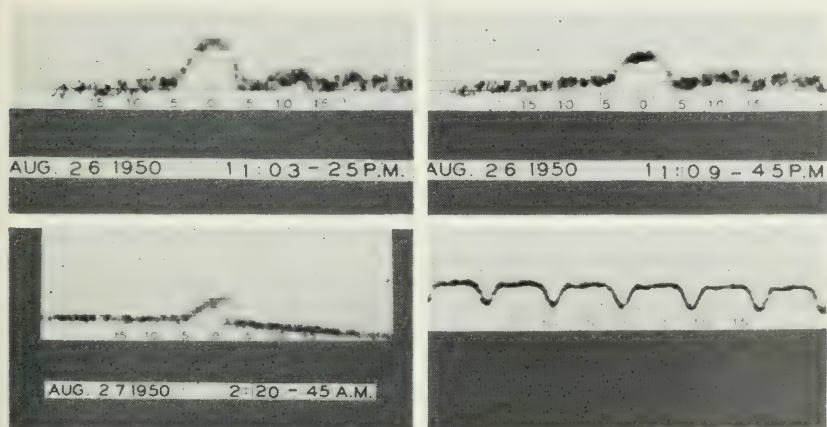


Fig. 8—Received pulses and calibrating pulses.

the output signal would be at maximum amplitude due to addition of energy coming from the two paths. This amplitude is, of course, twice that of the signal from one branch alone. In the experiment the plunger in one branch was left fixed and the attenuator in that branch left set at zero. The path through this branch then represented the normal transmission path for an actual system. The path through the other branch could be made to correspond to spurious paths having different amounts of delay and attenuation simply by adjusting the position of the reflecting plunger and the setting of the attenuator. A series of photographs were taken of pulses resulting from these different amounts of delay and attenuation.

The first three pictures of Fig. 10 were taken with the path lengths exactly equal. When the amplitudes were also equal there was complete cancellation. As the signal in one branch was attenuated the amplitude of the resultant pulse increased until it became equal to that of the original pulse as shown in the third picture. Increasing one path by one-half wavelength brought the signals from the two branches into phase and they added up to double amplitude as seen in the fourth picture. It should be pointed out that although in our experiment we changed delay by 0.36 millimicroseconds in going from the first minimum to the first maximum, in free space a change of delay of only 0.125 millimicroseconds

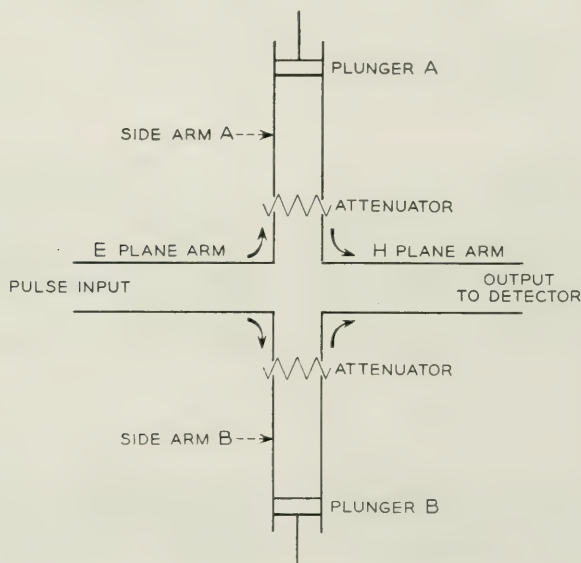


Fig. 9—Apparatus to simulate two-path transmission.

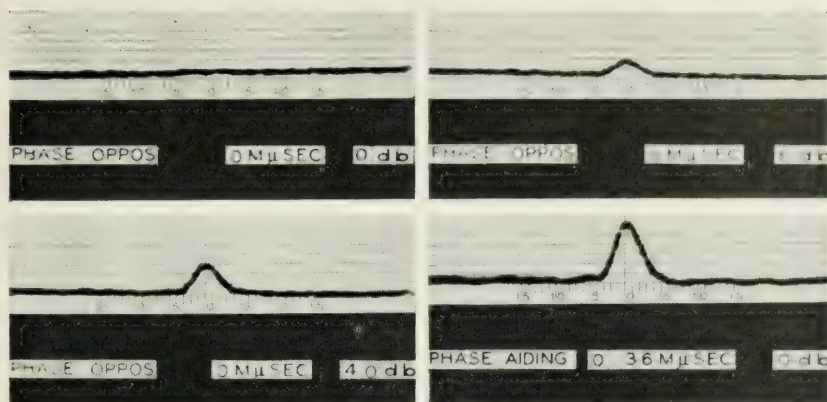


Fig. 10—Simulated two-path transmission.

would be required. The discrepancy lies in the large ratio between the phase velocity and group velocity in the wave guide used whereas in free space this ratio is, of course, equal to unity. In free space the amount of delay required to go from a maximum to a minimum signal corresponds to a change of path difference of only about one and one-half inches. With only this slight shift required to change conditions from those shown by the first picture of Fig. 10 to those shown by the last, it is not at all surprising that the received signal is very unstable during time of multi-path transmission.

Fig. 11 shows the effect of changing relative phases in 90° steps while keeping the amplitudes equal to each other. It is seen that even with the carriers in direct opposition cancellation is far from complete due to the

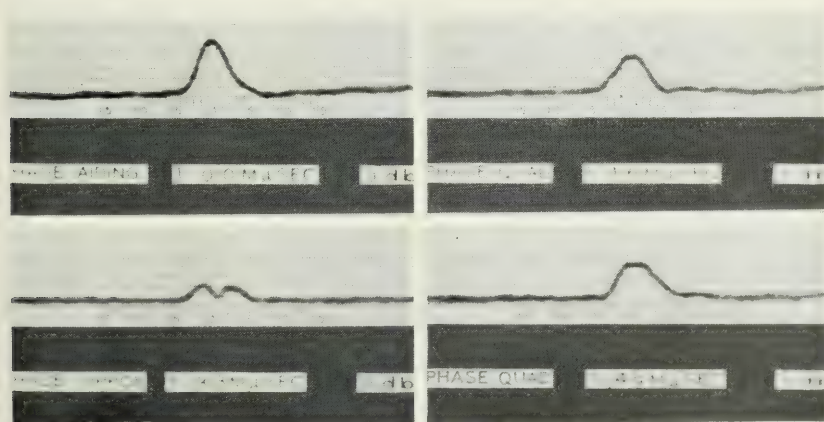


Fig. 11—Simulated two-path transmission.

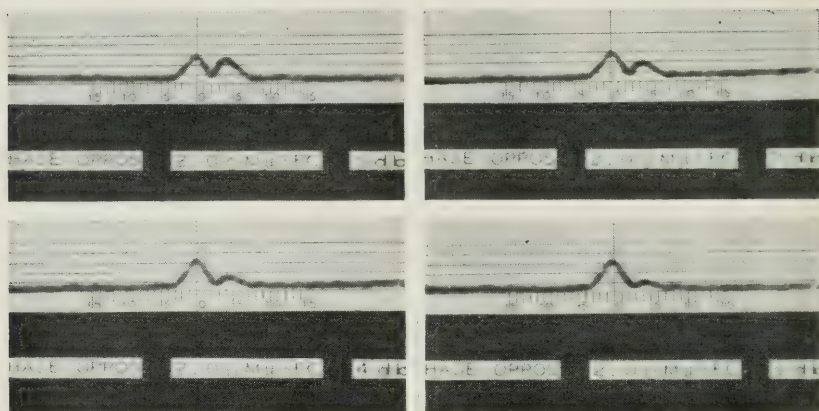


Fig. 12—Simulated two-path transmission.

relative delay between the two component pulses. Flat topped pulses seem to be characteristic of conditions where the two carriers are in phase quadrature and about equal in amplitude.

Fig. 12 shows a set of conditions with a constant delay difference of 2 millimicroseconds (corresponding to a path difference in free space of about 2 feet). For the pulses shown on Fig. 13 there was a constant delay difference of 7.34 millimicroseconds, enough to provide complete separation of the pulses. The carriers were in phase opposition but with this amount of separation there is no overlap of pulses and the results would have been the same if the phase had been aiding. Any increase of delay

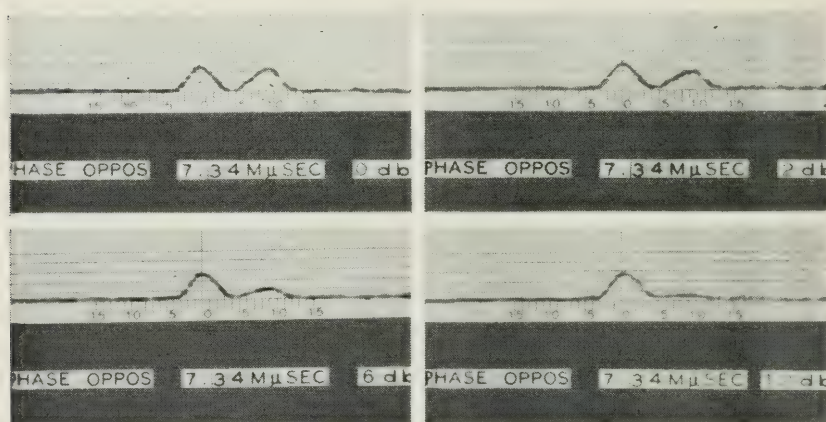


Fig. 13—Simulated two-path transmission.

beyond this point results only in moving the pulses farther apart and has no effect upon pulse shape or amplitude.

The experimental set up just described proved to be somewhat unsatisfactory since it was not possible with it to produce phase opposition between the two carriers without having zero delay difference between the two paths or a difference of at least 0.66 millimicroseconds. For the length of pulse used this latter amount of delay difference is sufficient to prevent anything like complete cancellation of pulses. In fact the amplitudes of the two resultant peaks are only about 12 db below the peak amplitude of the original pulse. From this we know that for the natural path any fade which appeared to be complete must have resulted from path differences of less than 0.66 feet, in fact from differences of less than about one-half foot.

SUMMARY

The pulse experiment results indicate that over one particular path at least there is, at times, transmission of microwaves by at least two, and probably more than two, paths. Path differences involved are from a fraction of a foot up to about seven feet, differences of less than about three feet being the most common. These results agree with those obtained by other methods. These multi-path effects result in bad distortion of very short pulses and even in the presence of entirely separate spurious pulses. These effects put a definite lower limit on pulselength and spacing between pulses in a pulse transmission system. The limit depends upon the amount of distortion which can be tolerated and also upon the percentage of time such distortion can be accepted. No statistical data were recorded.

With the laboratory equipment for simulating transmission over two paths, many of the waveforms obtained over the natural path could be duplicated. There were times, however, when the waveforms received by way of the natural path were too complicated to be explained by transmission by as few as two paths.

ACKNOWLEDGEMENTS

Space does not allow giving credit to all of the many people who contributed to the success of this project. A. F. Dietrich made the mechanical layouts for most of the equipment and supervised its construction as well as assisting in taking data and in many other ways. I also wish to thank J. C. Schelleng and W. M. Goodall for guidance and suggestions. G. M. Eberhardt furnished the traveling-wave-amplifier circuits.

Properties of Ionic Bombarded Silicon

BY RUSSELL S. OHL

(Manuscript received August 23, 1951)

This paper deals with a new and very interesting technique by which the properties of silicon surfaces are altered very materially by bombardment with ions of such gases as hydrogen, helium, nitrogen and argon. The change in rectifying properties has been of special interest but there have been considered also changes in the structural features of the material itself. The effects of bombardment on the rectifying properties are illustrated by a series of characteristic curves systematically arranged to bring out the effects of the several variables of experiment such, for example, as ion velocity, intensity of bombarding current, length of time of bombardment, kind of gas, and the temperature of the specimen during bombardment. The effect of bombardment on materials contaminated with impurities is also illustrated. It is of particular practical importance that silicon contaminated with boron to the point where it shows relatively little rectification can be modified by bombardment to make it even better than most unbombarded materials.

Some years ago, the writer discovered that the electrical properties of silicon surfaces could be greatly modified by bombardment with positive ions. The ions in question were generated in a low pressure discharge in some gas, like hydrogen, helium or nitrogen, and after passing through a perforated cathode were accelerated to a suitable velocity before impinging on the surface to be treated. This scheme may be contrasted with other methods subsequently reported for treating germanium¹ in which high-velocity ions were derived from radioactive sources. Preliminary results of the present research were described in a paper entitled *Silicon Transistors*, by W. J. Pietenpol and the writer, presented at an Electronics Conference held at the University of Michigan, June 22, 1950. Since that time exploration has continued with a view both to learning about basic principles and about possible practical applications.

Editorial Note—Since the resurgence of interest in point-contact rectifiers, considerable research has been carried on into the characteristics of silicon and germanium. The author of this paper was a pioneer in this new field of study, as evidenced, for example, by Patent No. 2,378,944, applied for on July 26, 1939, and Patent No. 2,402,839, applied for on March 27, 1941. More recent work has been described in a large number of text books and technical papers such as *Electrons and Holes in Semi-Conductors* by William Shockley, D. Van Nostrand, 1950, and numerous papers by Lark-Horowitz published mostly in *Physical Review*. The work described in the accompanying paper is a continuation of this long research.

¹ Brattain and Pearson, *Phys. Rev.*, **80**, Dec. 1950.

The present paper gives the results of some more recent experiments made with improved equipment. Also described briefly are some related experiments in which silicon is bombarded with alpha particles derived from radioactive polonium. The overall results of this work indicate rather clearly that with suitable variations of bombarding voltage, target temperature and time of exposure as well as impurity content in the base material, it is possible to prepare to specification silicon surfaces having a wide range of properties. From the materials so treated it has been possible to construct improved forms of signal rectifiers, harmonic generators, transistors, modulators, gating devices and also photo-electric cells. It is particularly significant that the voltage range over which these newer devices can be operated has been greatly extended, thus making them useful in places not previously regarded as possible. Since these new surfaces appear to be readily reproducible in large numbers and since they are electrically tough, chemically stable and show no unsatisfactory temperature or aging effects, it would appear that bombarding techniques should have considerable practical value.

This paper is concerned mainly with the practical aspect of ion bombardment of silicon, namely its effect on the voltage current characteristics at low frequencies. Equally important, perhaps, are its theoretical aspects, particularly with regard to the interpretation of the rather pronounced changes in the properties in light of presently-accepted views of solid-state physics. These aspects are not covered in this paper.

METHOD

The bombardment process referred to above consists of exposing the silicon surface to ions that have previously been accelerated to energies in the range from about 100 electron volts to about 30 kilo-electron-volts. A recent setup is illustrated in Fig. 1. The electrons from a tungsten cathode are accelerated toward a grid which is at a positive potential with respect to the cathode. Many of the electrons pass a short distance beyond the grid and return for ultimate capture. Ionization due mainly to the impacts of electrons with gas molecules takes place in this turn-around region, producing amongst other things positive particles. Electrodes are so proportioned that this ionization is fairly uniform over the grid area.

The silicon specimen to be bombarded is made negative with respect to the filament. This accelerates the positively charged particles toward the target. The latter rests on a graphite plate heated by a coil below, carrying high-frequency currents. A thermocouple with suitable connections to the exterior makes possible an adequate measurement of

temperature. The apparatus will accommodate circular surfaces as large as $1\frac{1}{2}$ inches diameter. The gases from which ions are derived are admitted through the gas inlet. Thus far experiments have been made with hydrogen, helium, nitrogen and argon. The bombarding voltages have as already noted, been varied from 100 to 30,000 volts and the surface temperatures have been varied from about 20°C to 400°C . The effects of these several variables will be discussed more fully below.

SAMPLE PREPARATION

The material to be bombarded is usually prepared in batches of about 300 grams in fused silica crucibles roughly cylindrical in shape.² After solidifying, the cast is ground to a cylinder approximately $1\frac{1}{2}$ inches

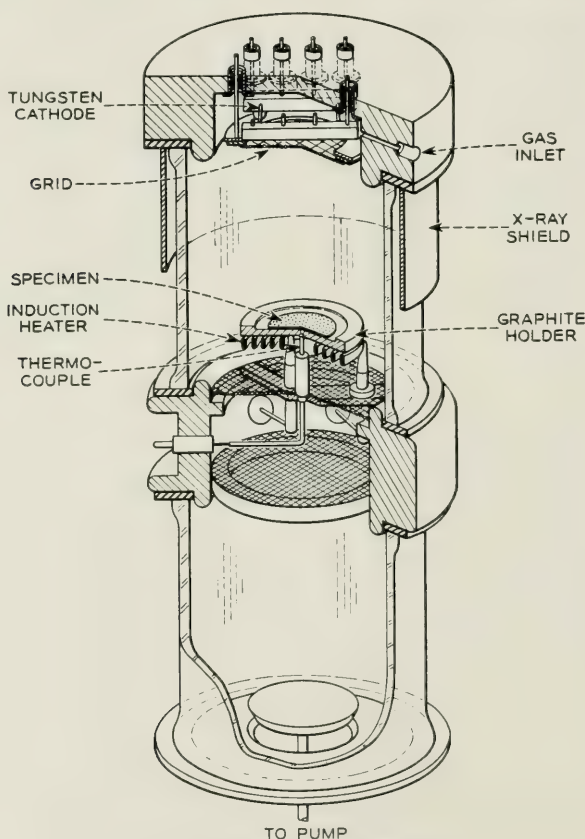


Fig. 1—Bombardment apparatus.

² Scaff, Theuerer, and Schumacher, A.I.M.M.E., **185**, pp. 382-392, 1949.

diameter. This has the effect of removing some of the contaminating impurities derived from the crucible as well as providing samples of convenient size. This $1\frac{1}{2}$ -inch cylinder is then sliced transversely into thin wafers which subsequently are polished on one side. Except as otherwise noted the material covered by this paper had an impurity content of about 0.1 per cent. The exception will be found in the data of column (a) of Fig. 8.

BOMBARDMENT PROCEDURE

The wafer, as prepared above, is placed in the bombarding apparatus with the rough face contacting the graphite support. The vacuum chamber is sealed by placing the ion generator in position and the whole assembly is evacuated. The sample is then heated to the proper temperature and the desired kind of gas is admitted, the pressure being estimated from the ion current. When stable conditions prevail, the accelerating voltage is applied to the target and the bombardment is carried out for the proper length of time. A convenient current density is 5 microamperes per square centimeter of target area. The target area of our present apparatus, including the silicon and a portion of the graphite support, being 20 square centimeters, the ion current is generally around 100 microamperes. The dosage is sometimes specified in microcoulombs.

After bombardment, the sample is removed from the apparatus and the rough surface is covered with a thin layer of evaporated rhodium. For most of the tests outlined below the $1\frac{1}{2}$ -inch diameter wafers were cut into $\frac{1}{8}$ -inch squares, a size convenient for testing.

GRAPHICAL REPRESENTATION

In considering the merits of non-linear materials such as silicon, perhaps the simplest and most useful characteristic is the voltage-current relation. If this is plotted to a linear scale, it results in a smooth curve of the general form shown in Fig. 2a. Specific curves obtained in practice may depart widely from that shown but in general, all may be regarded as made up of two semiparabolas, one in the first quadrant and one in the third, joined by a nearly horizontal straight line. For present purposes, we shall further simplify this idealized characteristic by considering it as made up of three straight lines. The first, AB, is associated with the reverse voltage current characteristic. The third, CD, is associated with the forward voltage current characteristic. These two characteristics are joined by the nearly horizontal line, BC. The slopes of these three lines correspond to resistances. The section BC for example, corre-

sponds to a region in which resistance is very high. The points B and C are particularly important for they represent points of inflection where the resistance undergoes rapid change and the material is departing most markedly from Ohm's Law. Ideally they should be sharp but in practice there is usually considerable curvature. Though either inflection point could presumably be used in detection processes, the point to the right of the origin is for practical reasons, usually preferred. Point B defines a voltage E_B at which substantial backward currents flow. It is referred to simply as the *reverse voltage*. In a similar way, point C defines a *forward voltage* E_F . The distance between B and C ($E_B + E_F$) will be referred to as the *inflection interval*. The difference in these quantities ($E_B - E_F$) is also of interest. One-half of this voltage difference is referred to as the *self-biasing voltage*. It is a significant quantity readily measured in practice by noting the d-c voltage across a large condenser placed in series with the crystal and a supply of 60 cycles AC. For detectors, point C should preferably be close to the origin and E_F should

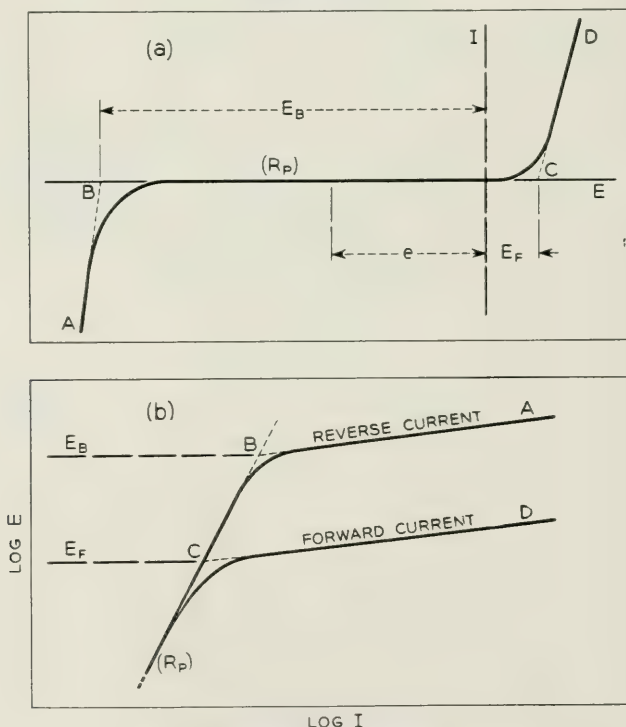


Fig. 2—Idealized characteristic curves.

be small. For certain kinds of voltage limiters, E_F should be large. In either case the inflection interval should be large.

In an alternate graphical representation, see Fig. 2b, voltage-current data are plotted to a log-log scale. This form of representation is of value in determining the resistance (R_p) at small voltages. Corresponding points on the two curves shown in Fig. 2 are identifiable by the letters A, B, C, and D. Curves of both kinds are used interchangeably to show the effects of the several variables of the experiment.

EFFECT OF CONTACT PRESSURE

In point contact rectifiers,³ pressure is of considerable importance. Usually the best pressure is a compromise between good electrical characteristics, usually obtainable only with light pressures, and good stability usually obtainable with higher pressures. Experiments have been performed with a range of contact pressures both on bombarded and unbombarded materials. In general, the results are highly variable, particularly in the case of unbombarded material. From this wide range of data, however, two characteristics have been selected that may be regarded as typical for 10-gram and 60-gram pressure. They are shown in Fig. 3 for silicon taken from nearby portions of the same sample. Significant points on these several curves may be compared with their idealized counterparts shown in Fig. 2. Although the samples chosen show somewhat more than the usual intrinsic resistance typical of p-type silicon, the effects of contact pressure are nevertheless regarded as representative. As indicated in Figs. 3a and 3b, the effect of increased contact pressure,⁴ particularly in the case of unbombarded material, is of reducing the low voltage resistance, R_p , see Fig. 2b. The more desirable higher resistance is obtainable only with light contact, a condition unfavorable for high mechanical stability. In the case of bombarded material, the effect of contact pressure is less important. Thus it is possible in this case to incorporate in the design higher contact pressures and obtain thereby higher stabilities. For purposes of this paper a contact force of 10 grams has been accepted as standard.

In addition to showing the effect of contact pressure, Fig. 3 shows some overall effects of bombardment. It will be noted, for example, that the effect of bombardment, see Fig. 3b, has been that of shifting the plots of Fig. 3a to the left by several orders of magnitude. Thus the resistance (R_p) is increased by a factor of more than 10,000. It is to be noted also

³ Scaff and Ohl, *Bell System Tech. J.*, **26**, Jan. 1947.

⁴ Really contact force.

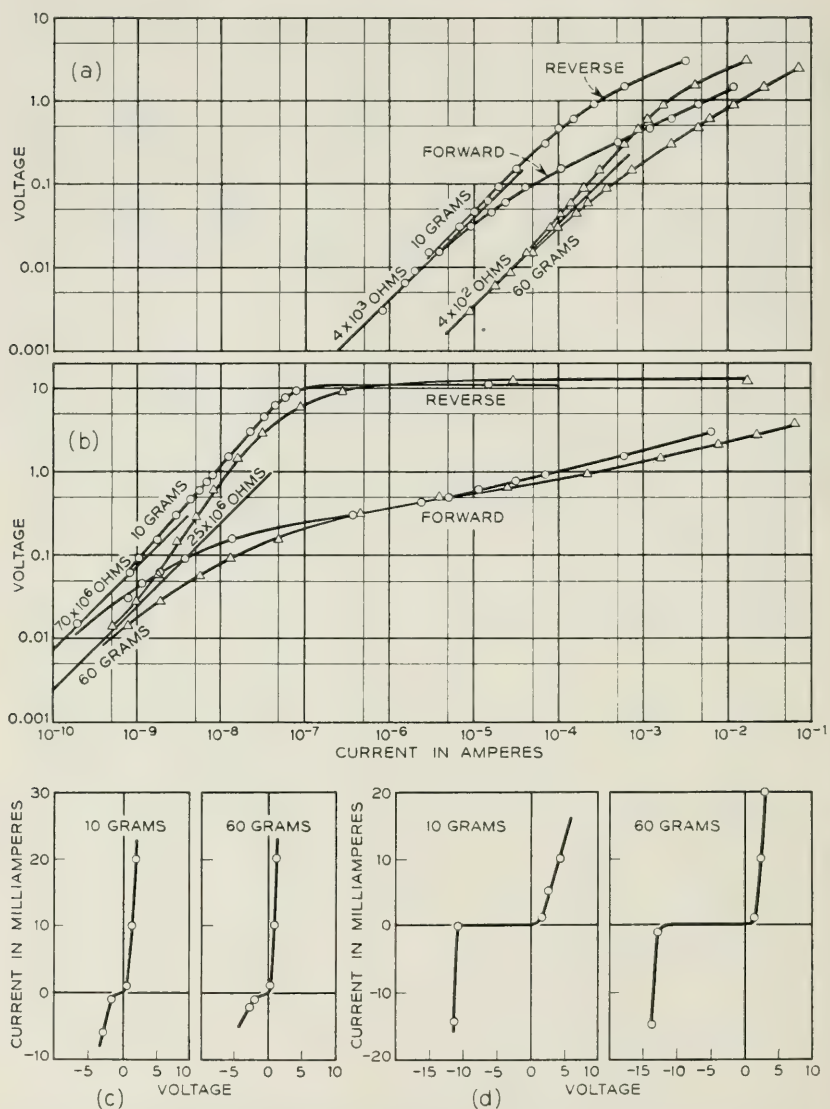


Fig. 3—Characteristic curves. (a) and (c) unbombarded silicon, (b) and (d) silicon bombarded with 30-kv helium ions.

from a comparison of Fig. 3a with Fig. 3b that at the one volt level, the ratio of forward to reverse currents for the unbombarded case is about twenty, whereas that for the bombarded case, is more than 10,000. At other levels the difference is even greater. Referring particularly to Figs. 3c and 3d, it will be seen that one effect of bombardment is that of separating the two significant points of inflection B and C. That is, the inflection interval has been notably increased. This increase is the result of a small increase in the forward voltage and a very substantial increase in the reverse voltage.

EFFECT OF TYPE OF GAS

Four high purity gases were tested as ion sources, namely, hydrogen, helium, nitrogen and argon, having atomic weights respectively of 1, 4,

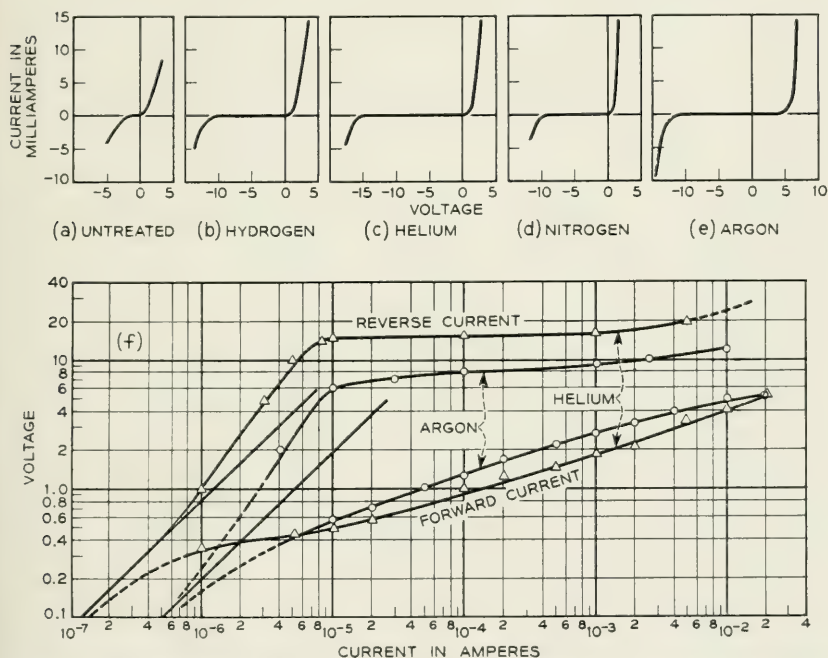


Fig. 4—Characteristics showing effect of various gases all with a bombarding potential of 30 kv.

14 and 40. While all four gases worked very well, helium was the easiest to handle. In the course of the tests, identically prepared samples each $\frac{1}{8}$ -inch square, taken from a high-purity silicon melt, were bombarded with ions formed in the particular gas under test. Particular conditions known to be good for producing good rectifier units were adopted as standard for these tests. They corresponded to a total bombarding charge of 600 microcoulombs per sq. cm., a surface temperature of 395°C a contact force of about 10 grams and a bombarding potential of 30 kv.

That the effect of bombardment varies with different gases is seen at a glance from the characteristic current-voltage curves shown in Fig. 4. Figs. 4b to 4e in particular indicate that as compared with an untreated sample, Fig. 4a, the effect of bombardment is in general that already noted of separating the two significant points of inflection, B and C. A rather substantial increase in the forward voltage appears in the case of argon as compared with hydrogen, helium and nitrogen. In contrast with a small increase in the forward voltage resulting from the bombardment of helium, there is a very substantial increase in the backward voltage. Though substantial for all four gases, the effect of bombardment is largest for helium with progressively smaller effects noted respectively for argon, hydrogen and nitrogen. A particular characteristic of helium bombardment, as compared with argon, not readily appreciated from a linear scale, is shown in Fig. 4f. It will be noted that at the one volt level, the ratio of forward to reverse current for helium is about 130 whereas for argon it is about 25. At other levels the difference is even greater. At the moment helium is regarded as a preferred source of ions.

The log-log current-voltage curves show as before that the lowest voltage at which substantial forward currents flow occurs in helium, while the highest forward voltages occur for argon. In a similar way the voltages for substantial reverse currents are highest for helium and lowest for argon. The sharp break in the reverse current characteristic, evident in these cases, has been observed so generally that it is now accepted as typical of bombarded surfaces.

EFFECTS OF TEMPERATURE

Investigations have been made of the properties of silicon surfaces as affected by the temperature at which bombardment was carried out. This has been done not only for surfaces used as rectifiers but surfaces used as transistors and as photo-electric cells as well. In the case of rectifiers, a procedure was adopted similar to that used in the previous tests. Measurements were made at five different temperatures ranging from

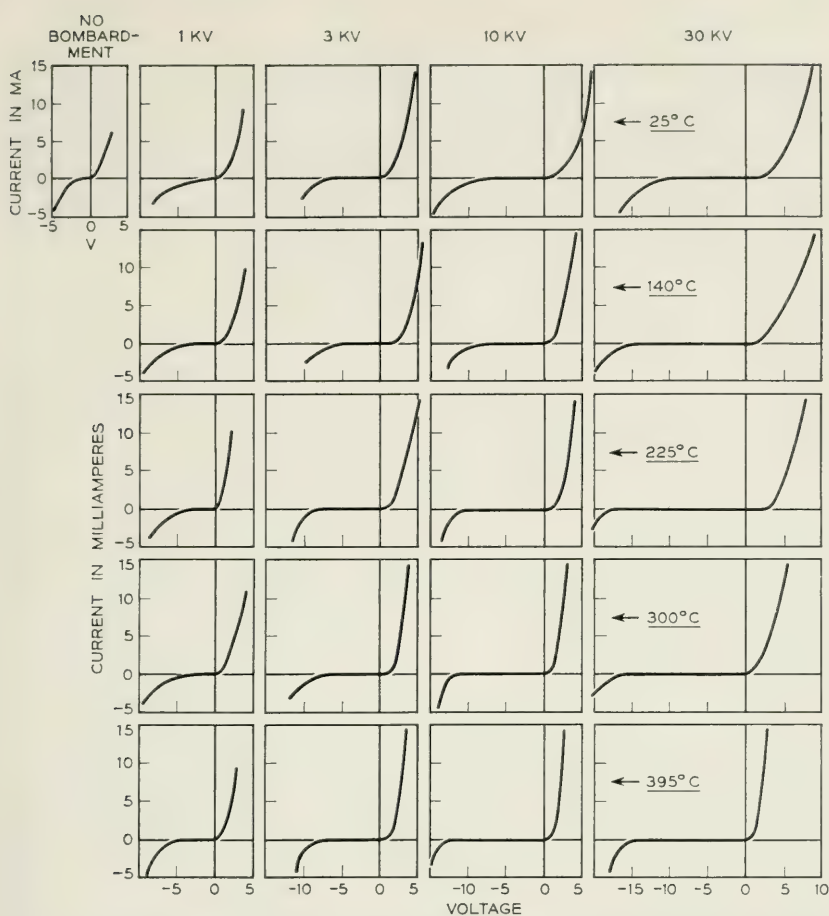


Fig. 5—Characteristics showing effects of voltage and temperature variation.

25°C to 395°C each at accelerating voltages of 1 kv, 3 kv, 10 kv and 30 kv using helium gas. The data so obtained were useful not only for studying the effect of temperature but useful in the studies of the effect of ion velocity as well. The latter will be discussed in the following section.

The results of the above measurements are plotted in Fig. 5. They are further summarized in Fig. 6a. The latter figure, in particular, indicates that as rectifiers, there is little choice of surface temperature between about 250°C and 400°C. It has been found, however, that for temperatures below about 250°C the point contact seems to be more vulnerable to electrical shock.

EFFECT OF ION VELOCITY

The effect of ion velocity (bombarding voltage) has been investigated for several types of silicon. The effects vary with the different types. Typical results are those given in Fig. 5 already referred to. It will be noted from a comparison of the data for a particular temperature, say 300°C , that the principal effect of increased ion velocity is that of increasing the reverse voltage. Values of these reverse voltages E_B and also the

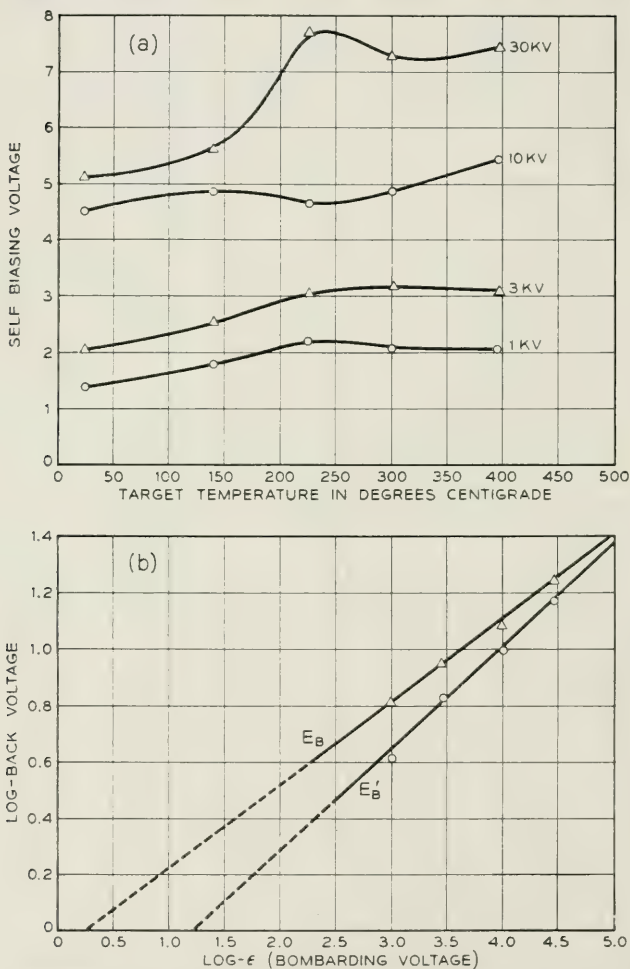


Fig. 6—Summary of data of Fig. 5. (a) effect of temperature and bombarding voltage on self biasing voltage; (b) effect of bombarding voltage on reverse voltage.

TABLE I—EFFECT OF BOMBARDMENT VOLTAGE ON E_B AND E_F

Surface Temp. Deg. C.	30 KV		10 KV		3 KV		1 KV		No Bombardment	
	E_B	E_F	E_B	E_F	E_B	E_F	E_B	E_F	E_B	E_F
395	16.7	1.3	14.0	1.5	10.0	1.5	7.1	1.0		
300	16.5	1.5	12.5	1.2	10.0	2.0	6.5	1.0		
225	18.5	3.6	12.5	1.5	9.6	1.5	5.5	0.3		
140	17.5	2.5	9.8	1.5	6.7	2.5	7.0	1.2		
Mean	17.3	2.0	12.1	1.4	9.1	1.9	6.5	0.9	2.4	0.5
Mean E'_B	14.9		9.7		6.7		4.1			

corresponding forward voltages E_F have been scaled from the above drawings and have been tabulated below. Since they vary only slightly with surface temperature, only their mean values are regarded as significant. Mean values of E_B are plotted in Fig. 6b.

In order to isolate further the effects of bombardment we have subtracted from the mean values of E_B value of E_B for untreated silicon. Thus the curve marked E'_B represents the improvement in backward voltage that has accrued from bombardment alone. This is also tabulated as the mean E'_B in Table I.

EFFECT OF TOTAL CHARGE

Tests have been made to determine the effect of time of bombardment on the rectifying properties of silicon surfaces. In these tests, specimens taken from neighboring regions of the same melt were exposed for progressively longer periods all at the same bombarding potential of 30 kv and the same rate and density of application, 5 microamperes per square centimeter. Representative current-voltage characteristics are plotted in Figs. 7a and 7b for two neighboring regions. The results are summarized in Fig. 7c. The latter show a rather rapid improvement of back voltage E_B with total charge up to perhaps 50 microcoulombs per square centimeter.⁵ Thereafter the improvement is small. For purposes of comparison, there is plotted as a vertical line a value of bombarding charge that would account theoretically for one positive ion in each unit crystallographic cell on the surface layer. This suggests that when all surface cells have been penetrated by a single ion, no marked increase in back voltage can be effected.

⁵ Specifying results in terms of microcoulombs implies that bombarding effects are independent of the rate of application. This is known to be true only between factor limits of $\frac{1}{2}$ and 2 of the bombarding current.

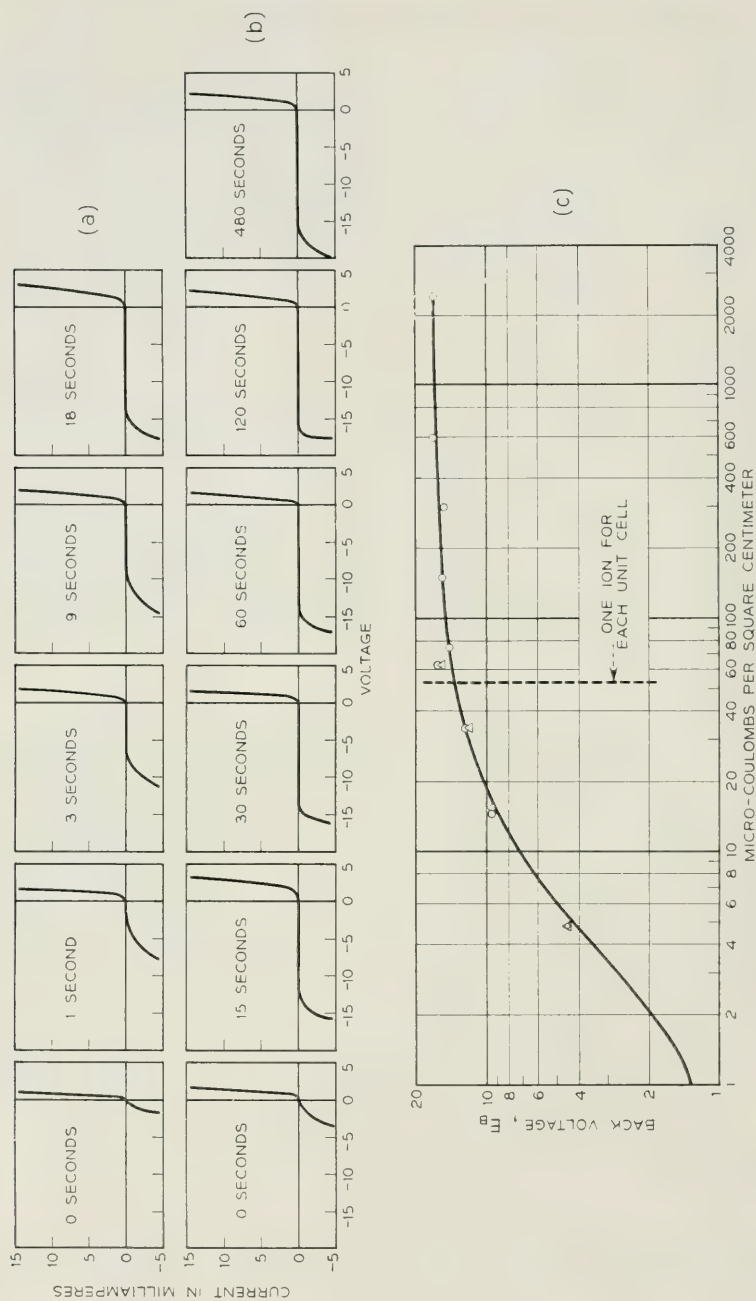


Fig. 7—Effect of time and bombarding charge on back voltage.

EFFECT OF MATERIAL COMPOSITION

Thus far discussions have centered around a single type of high purity material that was regarded as representative. It is of interest to examine the effect on other materials particularly those in which impurities have been added. For this purpose tests were made on comparable samples from four sources all bombarded for two minutes with 5 microamperes of current and each with five representative bombarding voltages. The results are illustrated by the curves shown in Fig. 8. The four columns correspond to progressively higher percentages of impurities beginning with (a) on the left as a material having an impurity content believed to be less than 0.01 per cent. The impurity content of (b) is not known accurately except that it lies between (a) and (c). The material represented by column (c) was produced by adding 0.02 per cent boron⁶ to a material illustrated in column (a). The last column (d) was produced by adding 0.1 per cent boron to the material illustrated in column (b).

It is to be noted from Fig. 8 that marked changes in the voltage-current characteristic may be effected by bombardment for all degrees of the impurity content shown. It is especially interesting that in columns (c) and (d) corresponding to materials contaminated with boron to the point where nonlinearity is almost absent, rectification can not only be restored but indeed the product may be made better than the best unbombarded material.

EFFECTS OF ALPHA-PARTICLE BOMBARDMENT

The close relationship between helium ions such as generated above and alpha particles such as emanate from radioactive materials suggests that the latter may be used for the bombardment of silicon surfaces. A few experiments of this kind have been made with results that are not only interesting but possibly useful. For these tests, four sources of alpha particles were obtained. They consisted of $\frac{5}{16}$ inch square pieces of nickel on which had been plated a thin coating of polonium followed by a covering of gold. The initial strength was 4 millicuries per square centimeter. The half-life of polonium is 140 days.

The process of bombardment consisted simply of placing the polished surface of a standard silicon square against the layer of gold and examining the same periodically. Tests of four samples were carried out simultaneously. The results are given in Table II. The data for Sample No. 1 departs so markedly from the mean that it may be disregarded. Since

⁶ Boron is a particularly active agent in effecting changes in the properties of silicon.

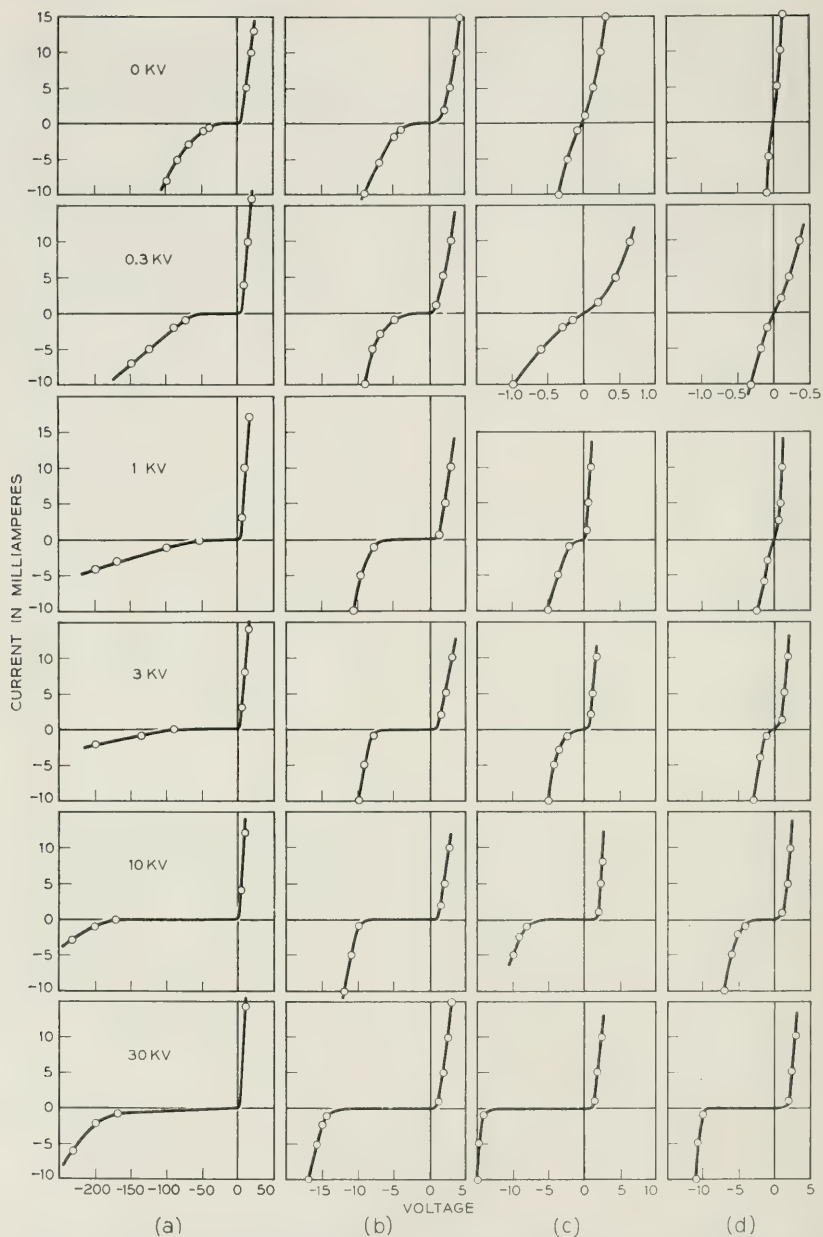


Fig. 8—Effect of impurity content. (a) hyper-purity silicon, (b) high-purity silicon, (c) hyper-purity silicon plus 0.02 per cent boron, (d) high-purity silicon plus 0.1 per cent boron.

the data for self-bias are the result of a direct measurement while those for forward and reverse voltage are transcribed from a cathode-ray plot, they are perhaps the most significant. Samples 1 to 3 represent specimens of increasing degrees of purity.

These alpha-particle bombardment experiments indicate rather definitely that results may be obtained similar to those obtained from bombardment with gaseous ions and, like the ion bombardment, they tend to produce high resistance surfaces.

TABLE II

Days of Exposure		Sample Number			
		1	2	3a	3b
4	Self Bias-Volts	<1	6	180	60
4	Reverse Voltage	<1	20	380	130
4	Forward Voltage	<0.5	<0.5	<1	<1
13	Self Bias-Volts	<1	160	200	190
13	Reverse Voltage	1	360	440	420
13	Forward Voltage	<0.5	<0.5	<1	<1
39	Self Bias-Volts	<1	60	130	100
39	Reverse Voltage	0.5	480	500	500
39	Forward Voltage	<0.2	350	180	180
56	Self Bias-Volts	<1	70	150	110
56	Reverse Voltage	0.5	440	560	560
56	Forward Voltage	0.3	320	320	240
66	Self Bias-Volts			100	85
66	Reverse Voltage			560	520
66	Forward Voltage			200	320

MECHANICAL EFFECTS OF BOMBARDMENT

The marked changes in the electrical properties of silicon imposed by bombardment strongly suggest that bombardment may also impose a corresponding change in the lattice structure and that this might be detected by suitable optical methods. Attempts were made at an early date to detect such changes. To this end a mask of nichrome ribbon 5 mils wide and 1 mil thick was laid over a sample of silicon during bombardment. An optical examination of the surface showed that after bombardment in the case of helium the surface on either side of the mask was elevated whereas in the case of argon it was depressed. This result has since been confirmed by one of the authors's colleagues, Dr. F. W. Reynolds, who has found that in cases of prolonged bombardment by helium the adjacent surfaces may be elevated by as much as 225 Angstroms⁷ while in the cases of prolonged bombardment by

⁷ One Angstrom is 10^{-8} cm.

argon the adjacent surfaces may be depressed by as much as 130 Angstroms. Further investigation of this phenomenon is under way.

STABILITY OF BOMBARDED SURFACES

No extended test has yet been made of the stability of bombarded surfaces but results extending over more than two years are encouraging. Similarly, rectifiers for the millimeter wavelength range, mounted without the usual protective impregnation, show little or no change at the end of a year.

In a few instances bombarded surfaces have been subjected to rather severe tests with results that suggest that under normal conditions they may be even more stable than surfaces activated by more conventional methods. For example, surfaces contaminated while cutting or while cementing them to their mountings have subsequently been cleaned with solvents such as alcohol and are substantially the same before and after treatment. In other cases, they have been heated in a flame to soldering temperatures with no appreciable effects. Even in the very severe case where the bombarded piece was heated to a cherry red and the superficially oxidized layer was removed with hydrofluoric acid the effects of bombardment were still evident. There was, however, considerable reduction in the tolerable reverse voltage.⁶ There is nothing in our experience to date to suggest that bombarded surfaces treated in accordance with the simple straightforward methods outlined above, are in any wise temporary in character.

CONCLUSIONS

The experiments reported above have shown that rather pronounced changes in the electrical properties of silicon may be produced merely by bombarding the polished surface with positive ions. The ratio of forward to reverse currents, for example, which for the usual untreated silicon is seldom more than a few hundred, can be made more than 10,000. Experiments show that the effect depends to some extent on the type of ion gas used, helium being a preferred medium. The effect depends also on the velocity of the bombarding particles, the total bombarding charge and to a lesser extent on the temperature of the specimen during bombardment. Good results are obtained from bombarding potentials of 30 kv with current densities of 5 microamperes per square centimeter for periods of one or two minutes. The temperature should preferably be about 300°C.

Ordinarily the properties of silicon are materially affected by impurity

content. In the case of bombarded silicon the effect is much less. More particularly it is possible to contaminate silicon with impurities such as boron to the point where its rectifying properties are almost completely lost and by bombardment it is possible to convert the crystal into a very useful rectifier. It is possible to produce results similar to the above by exposing the crystal to radioactive polonium. Bombarded materials appear to be relatively stable.

The writer wishes to express his appreciation of the encouragement and help of Dr. G. C. Southworth in the preparation of this paper, to A. J. Mohr, Jr., for his able assistance in the experimental work, and to numerous associates in Bell Telephone Laboratories for their assistance in preparing materials and in making special tests for which the author was not adequately equipped.

Mechanical Properties of Polymers at Ultrasonic Frequencies

BY WARREN. P. MASON AND H. J. McSKIMIN

(Manuscript received October 25, 1951)

Since the mechanical properties of solid polymer materials are largely dependent on the motions that segments of the polymer chains can undergo, to understand these properties one must use measuring techniques which can determine these motions. One of the most promising methods is to measure the reaction of polymer materials to longitudinal and shear waves over a frequency spectrum wide enough to determine the relaxation frequencies due to thermal motions of the principle elements of the chain. The presence of relaxations is indicated by a dispersion in the velocity and attenuation constants of the material, or a dispersion in the characteristic impedance of the material if the attenuation is too high to allow velocity measurements. A number of different types of measuring methods are described in this paper which make possible propagation and impedance measurements not only in solid polymers, but also in liquid polymers and in solutions of polymer molecules in typical solvents.

When these techniques are applied to long chain polymers in dilute solutions, the three relaxations observed correspond to motions occurring in isolated molecules since as the dilution increases, the molecules seldom touch. The lowest relaxation corresponds to a configurational relaxation of the molecule as a whole, the highest relaxation corresponds to the twisting of the shortest segment—containing about 40 repeating units—while the intermediate relaxation corresponds to a transient entanglement of chain segments. All three types of relaxations are present in pure polymer liquids but are spread out over a frequency range due to the perturbing effect of near neighbors of adjacent chains. The high frequency shortest chain relaxation can be traced in solid polymers of the linear chain type such as polyethylene and nylon and produces rubber-like response to mechanical shocks of very short duration.

I. INTRODUCTION

The mechanical properties of solid polymer materials are largely determined by what motions, parts or segments of the polymer chains can undergo. Toughness, mechanical impact strength and ultimate elongation depend on the facility with which the polymer molecule can be displaced.

If only a small motion of the polymer chain can occur within the time of the measurement, the material has high elastic stiffness coefficients and acts similar to a rigid solid. On the other hand, if significant segments of the polymer chain can move at the frequency of measurement, the elastic stiffness is much lower and rubber-like behavior results. An intermediate case, which occurs when the significant motion of the polymer molecule is near the relaxation time at the frequency of measurement, is that of a damping material such as butyl rubber. Even "long time" qualities of plastics such as creep, stress relaxation and recovery depend on the integrated displacements of rapidly oscillating segments of the chain.

One of the most promising methods for investigating these motions is to determine the reaction of mechanical waves on the polymer materials over a wide spectrum of wavelengths, eventually going to frequencies comparable with those of thermal vibrations of significant groups or segments in the macromolecules.

If one wishes to understand the origins of these motions it is necessary to measure the molecules in the form of liquids or solutions since then the segments of the molecule are less restrained by their neighbors and can perform all the possible vibrations. Polymer liquids are also interesting in themselves as sources of damping material. To apply these results to rubbers and solid materials, one then has to measure the modifications of the polymer chain motion caused by the close approach of near neighbors, by measuring the mechanical properties of these materials.

By using different types of techniques, these processes can be applied to molecules in solution, to liquid polymers and to solid polymers. The principal types of methods used for liquids are the torsional crystal, the torsional wave propagation system and the shear wave reflectance method, all of which are described in Section II. For solids an optical method and an ultrasonic method are described in Section V. All of these methods involve displacements of 10^{-6} cm or less so that non-linear effects are negligible.

All of these methods depend on setting up shear or longitudinal waves in the medium and observing either the velocity and attenuation of the wave, or the reaction of the medium back on the properties of the transducer. If the attenuation of a wave in the medium under consideration is low enough to permit the wave parameters, i.e., the velocity and attenuation per wavelength to be determined, the relaxation of some significant part of the polymer molecule is determined by the dispersion of the wave properties which occur, as shown by Fig. 1A, in the form of an increase in velocity and a maximum in the attenuation per

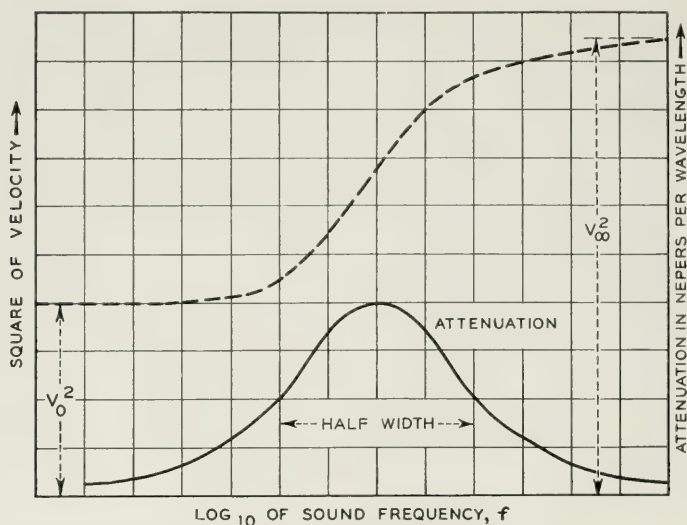


Fig. 1A—Velocity and attenuation for a medium with one relaxation frequency.

wavelength curve. If the variation in this relaxation mechanism is studied as a function of temperature and chain length, the type of segment may be determined. If, however, the attenuation of the medium is so high that its wave properties cannot be determined, some information can still be obtained by determining the loading, or mechanical impedance, that such a wave exerts on the driving crystal or transducer. If all the relaxations occur in the stress-strain relation, it can be shown that there is a reciprocal relation between the propagation constant $\Gamma = A + jB$, and the characteristic impedance per square centimeter Z_0 given by the equation

$$Z_0 \Gamma = (R + jX)(A + jB) = j\omega\rho \quad (1)$$

where A is the attenuation and B the phase shift per centimeter, R the mechanical resistance and X the mechanical reactance per square centimeter, ω is 2π times the frequency and ρ the density of the medium. A typical two relaxation mechanism¹ is shown by the curves of Fig. 1B. By assuming values for the stiffness and dissipation factors and fitting a theoretical curve to the measured values, the relaxation frequency or frequencies can be determined.

¹ All the relaxation mechanisms discussed in this paper are represented in terms of equivalent parallel electric circuits in which the resistance terms represent viscosities and the inverse of capacities represent shear elastic stiffnesses. In mechanical terms these correspond to a series of Maxwell models as discussed in a paper by Baker and Heiss to be published in the next issue.

The most information about the motions of isolated polymer chains can be obtained by investigating the properties of polymer solutions. This follows from the fact that in pure polymer liquids, and in solids, the mechanical properties are mainly determined by interactions between chains on account of the close packing of the chains. If, however, one dissolves the polymer molecules in a solvent, the inter-chain and intra-chain reactions can be separated as the dilution increases. When the polymer is in the order of one percent of the solvent, the chains on the average touch very seldom and the mechanical properties of the solution are determined by the properties of single molecules. As discussed in Section III, three types of chain segment motion have been isolated, (1) a configurational relaxation of the chain as a whole, (2) a position change of the shortest segment and (3) twisting of the shortest chain segment. Above the frequency of relaxation of this chain segment the joints of the polymer molecule become frozen and the chain becomes very stiff. These shortest chain relaxations occur also in pure polymer liquids, in rubbers and in non rigid solids with linear chain segments such as polyethylene. In pure liquids a lower frequency quasi-

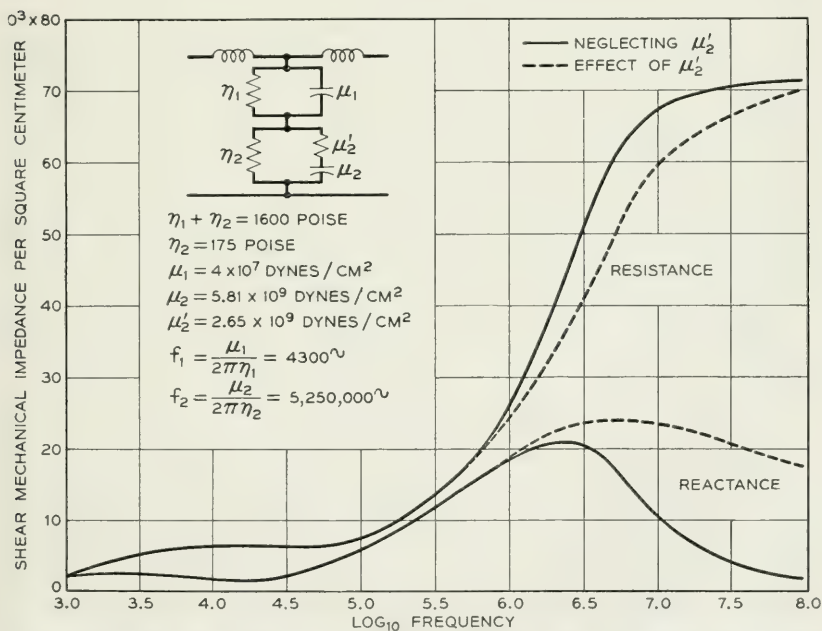


Fig. 1B—Mechanical impedance loading for a medium with two relaxation frequencies.

configurational relaxation also occurs for chain lengths greater than 60 elements but for chain lengths less than 40 elements this type of relaxation disappears. From the difference between the high frequency shear elasticities measured for polyethylene and nylon 6-6 and those measured for static pulls, it appears that there may be lower frequency relaxations in these materials as well.

II. METHODS OF MEASUREMENT FOR SOLUTIONS AND PURE POLYMER LIQUIDS

To measure the mechanical properties of such dilute solutions, shear waves have been used since for longitudinal waves the added stiffness caused by the dissolved polymers is very small compared to the stiffness of the solvent alone. The velocity and attenuation of a longitudinal wave are given by the equations

$$v = \sqrt{\frac{\lambda + 2\mu}{\rho}}; \quad A = \frac{2\pi^2 f^2}{\rho v^3} [\chi + 2\eta] \quad (2)$$

where λ and μ are the Lamé elastic constants, f the frequency, ρ the density, v the sound velocity, χ the compressional viscosity and η the shear viscosity. Since for a one percent solution of polyisobutylene in cyclohexane the shear elasticity does not exceed 90,000 dynes/cm², whereas the value of λ is in the order of 2×10^{10} dynes/cm², it is obvious that the longitudinal velocity would have to be measured to an accuracy of 1 part in 100,000 before the presence of polymer molecules could be ascertained. Attenuation measurements give some information on the added viscosity due to the chain molecules but since longitudinal attenuations are not easily measured below 1 megacycle, the most interesting frequency range is missed.

A pure shear wave in a viscous liquid is propagated according to the equation²

$$v = v_0 e^{-\sqrt{\frac{\pi f \rho}{\eta}} (1 + j) z} \quad (3)$$

where v is the transverse particle velocity, ρ the density, f the frequency, η the shear viscosity, $j = \sqrt{-1}$ and z the distance. For typical solvents, the attenuation is so high that wave motion cannot be measured. However the viscous wave produces an impedance loading on a crystal generating such a wave which can be measured by the change in the resonant frequency and the change in the resistance at resonance. The mechanical impedance per square centimeter caused by such a viscous

wave is equal to²

$$Z_0 = \sqrt{\pi f \eta \rho} (1 + j) = R_M + jX_M \quad (4)$$

This causes a change in resistance, and a change in frequency in a crystal generating a shear wave in the liquid equal to

$$\Delta R_E = K_1 R_M; \quad \Delta f = -K_2 X_M \quad (5)$$

where K_1 and K_2 are constants of the crystal which can be obtained approximately from the dimensions and piezoelectric constants of the crystal but which are more accurately obtained by calibration in known liquids. The constants K_1 and K_2 vary slightly with temperature and should be calibrated over a temperature range.

The first instrument to use a vibrational method for measuring viscosity was the vibrating wire method of Phillipoff.³ In this method a wire was vibrated in a liquid and the damping rate was used as a measure of the viscosity. Another method also applicable in the low frequency range is the transducer method of Ferry.⁴ In this method wires are vibrated by electromagnetic transducers and the resistance and reactance drag on the wires are measured by the change in the electrical resistance and reactance of the transducer. From the constants of the transducer, the equivalent viscosity and stiffness of the liquid can be measured.

In the medium frequency range a torsional crystal⁵ method was devised by one of the writers which has been applied in the frequency range from 10 to 150 kc. The torsional crystal is shown by Fig. 2. For these types of measurements the crystal usually is made of quartz with four electrodes of gold evaporated on the surface. Four wires are soldered on the surface and serve as supports as well as electrodes. The motion is all tangential to the surface and tests at Bell Laboratories and at the Franklin Institute,⁶ where a precision study of the torsional crystal has been made, have shown no observable longitudinal waves from the crystal surface. The process of measurement consists in measuring the

² W. P. Mason, *Piezoelectric Crystals and Their Application to Ultrasonics*, D. Van Nostrand, 1950, p. 340.

³ W. Phillipoff, *Physik. Zeits*, **35**, 1934, pp. 884-900.

⁴ T. L. Smith, J. D. Ferry and F. W. Schemp, "Measurement of the Mechanical Properties of Polymer Solutions by Electromagnetic Transducers," *J. App. Phys.*, **20**, No. 2, Feb. 1949, pp. 144-153.

⁵ W. P. Mason, "Measurement of the Viscosity and Shear Elasticity of Liquids by Means of a Torsionally Vibrating Crystal," *A.S.M.E.*, **69**, May 1947, pp. 359-367.

⁶ P. E. Rouse, Jr., E. D. Bailey, and J. A. Minkin, "Factors Affecting the Precision of Viscosity Measurements with the Torsional Crystal," Laboratories of the Franklin Inst., Report 2048, presented to Am. Petroleum Inst., May 4, 1950.

resonant frequency and resonant resistance of the crystal in a vacuum, then introducing the solution to be measured, the change in the resonant resistance ΔR_E and the change in resonant frequency Δf are determined by an electrical bridge. Several short cuts are possible if the mechanical impedance is not too high. By measuring the capacity at a frequency considerably higher than the crystal frequency, the resistance at resonance and Δf can be obtained by changing the frequency and resistance until a balance is obtained leaving the capacity unchanged. This method has been used to measure viscosity, and a recent precision study at the Franklin Institute⁶ has shown that it agrees with other methods to an accuracy of well under a per cent.

The torsional quartz crystal has been successfully used to measure liquids having a viscosity up to 10 poise, but above this viscosity the electrical resistance gets so high that it is hard to measure it since it is shunted by the much smaller reactance of the static capacitance of the crystal. A crystal of higher electromechanical coupling such as am-



Fig. 2—Cell and 80-ke crystal for shear viscosity and elasticity measurements of liquids.

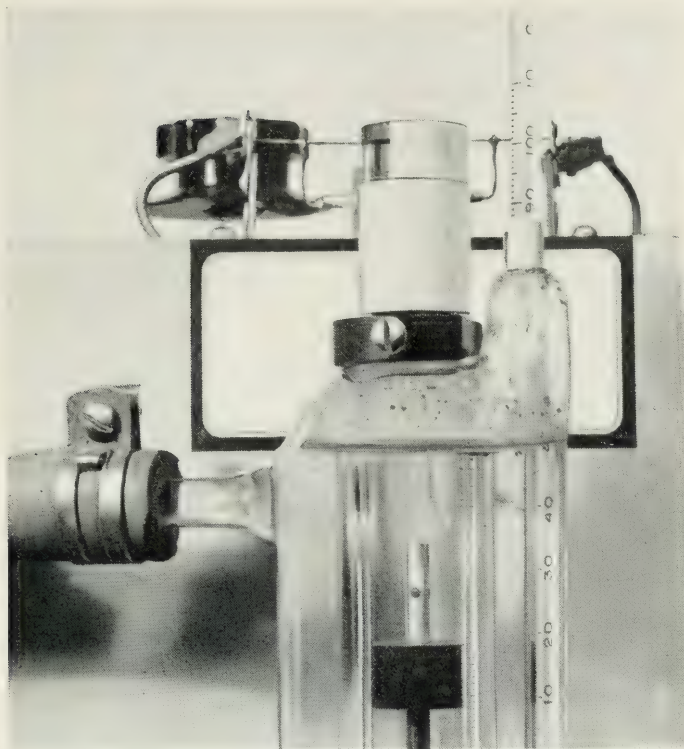


Fig. 3—Photograph of torsional crystal and rod.

monium dihydrogen phosphate (ADP) will cause the electrical resistance component to be smaller in comparison to the reactance of the static capacitance and hence can be used to measure higher viscosities. However, since wires cannot be soldered to the surface but must be glued, the crystal is much more fragile than quartz and its use has been abandoned in favor of another method which makes use of the phase and attenuation change in a torsional wave in a rod caused by the surrounding liquid whose properties are to be measured.

This method, devised by one of the writers,⁷ consists in sending a short train of torsional waves, periodically repeated, down a glass or metal rod. As shown by Fig. 3, the torsional wave is generated by a torsional quartz crystal soldered or glued to the end of the rod. These waves travel to the free end of the rod and are reflected back to the crystal where they are detected, amplified, and displayed on a cathode ray

⁷ This method is described by H. J. McSkimin, in a paper before the Acoustical Soc. of Am. in October, 1951.

oscilloscope. Echoes due to end to end reflections also appear, being attenuated by normal acoustic losses until they are undetectable by the time the next pulse is applied.

With only air surrounding the rod, a phase reference and amplitude reference are obtained for the first received wave (or subsequent echoes if greater sensitivity is desired). The rod is then immersed a definite length in the liquid to be measured, as shown in Fig. 4, with a resulting phase retardation and amplitude reduction. These are measured by employing the experimental circuit shown by Fig. 5. In order to allow the

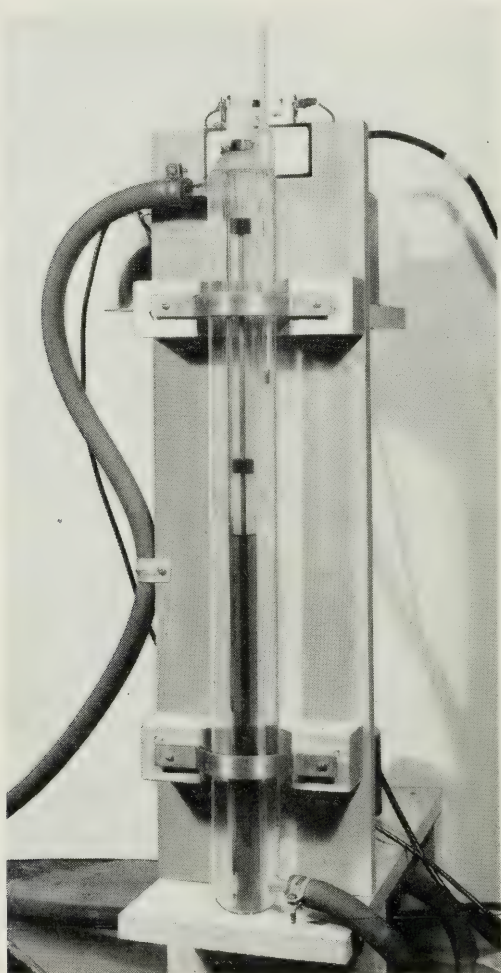


Fig. 4—Photograph of complete torsional wave measuring equipment.

use of one crystal for both receiving and transmitting, the crystal is put in the bridge circuit of Fig. 5 where a resistance and capacity are used to balance out the transmitted pulse so that it will not overload the amplifier. The relatively weak voltages generated by the incoming acoustic waves pass through directly. The gate circuit provides pulses of radio frequency voltage at repetition rates in the range of 20 to 100 per second with a synchronizing voltage supplied to the oscilloscope for the horizontal sweep. The frequency range of the device is from 20 to 200 kc. Both glass and nickel-iron rods were used, the latter having a very low frequency-temperature coefficient. With a 100-kc quartz

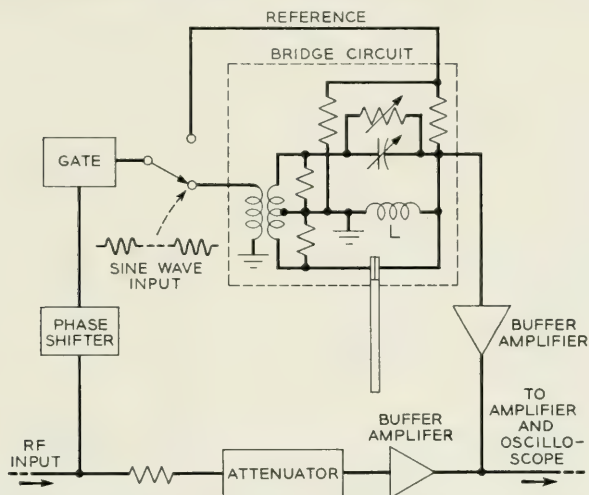


Fig. 5—Experimental pulsing circuit for measuring torsional impedance of liquids.

torsional crystal, a rod length of 21 inches and diameter of 0.2 inch were used. The entire crystal-rod assembly is placed inside a glass temperature control unit, as shown by Fig. 4, through which water can be circulated to provide temperatures in the range 0°C to 80°C . The test liquid is placed either directly into the inner bore of this water jacket, or in another tube which can be inserted from the bottom to surround the rod up to a fixed mark.

In use, both phase and attenuator settings were adjusted to balance the first received pulse against the continuous wave component passing through the attenuator. Cancellation for the duration of the pulse was visually indicated on the oscilloscope. A plot of balance phase and level is made as a function of the temperature. When the liquid is introduced an attenuation change ΔA and a phase change ΔB are required to

re-balance the circuit. These are measured by the amount of attenuation in nepers (1 neper = 8.68 db) and the number of radians phase shift required to re-establish a balance. An alternate method of measuring phase shift is to measure the change in frequency required to re-establish balance. If this method is used the phase shift change of the overall circuit with frequency has to be calibrated for the uncovered rod by

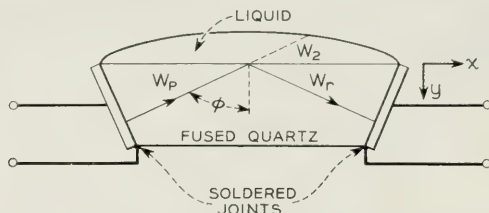


Fig. 6—High frequency shear reflection method for measuring shear impedances of liquids.

noting the frequencies for which 360° phase shifts (as measured by balance) occur in the circuit.

It is shown in the appendix that the torsional impedance of the liquid per square centimeter is given by

$$Z = \left(\frac{\rho v_0}{4} \right) \frac{a}{\ell} [\overline{\Delta A} + j\overline{\Delta B}] \quad (6)$$

where ρ is the density, and v_0 the sound velocity in the rod, a is the radius and ℓ the covered length of the rod and $\overline{\Delta A}$ and $\overline{\Delta B}$ are respectively the change in attenuation in nepers and the change in phase shift in radians to re-establish balance. If a very viscous liquid is used it may be necessary to correct for the fact that the torsional impedance may differ from the plane wave impedance as discussed in the appendix.

This device can measure liquids having dynamic viscosities from 10 poise to 1,000 poise with an accuracy of the order of 10 per cent. The frequency range covered may be from 20 to 200 kc depending on the size of the crystal used to drive the rod. Hence it supplements the torsional crystal method for very viscous liquids.

At frequencies above 500 kc, the torsional crystal becomes too small to be used practically and recourse is had to a high frequency pulsing method.⁸ As shown by Fig. 6, shear waves are set up in a fused quartz

⁸ W. P. Mason, W. O. Baker, H. J. McSkimin and J. H. Heiss, "Measurements of the Shear Elasticity and Viscosity of Liquids by Means of Ultrasonic Shear Waves," *Phys. Rev.*, **75**, No. 6, March 15, 1949, pp. 936-946. See also H. T. O'Neil, "Refraction and Reflection of Plane Shear Waves in Viscoelastic Media," *Phys. Rev.*, **75**, No. 6, March 15, 1949, pp. 928-936.

rod by means of a Y-cut or AT cut crystals soldered to a silver paste layer baked on the fused quartz surface. The particle motion of the shear wave is parallel to the large reflecting surface and hence only shear waves are reflected from this surface. These impinge on a second shear crystal which is connected to an amplifier and oscillograph. Since the attenuation in fused quartz is so low, a long series of reflected pulses appear on the oscillograph. When a liquid, whose shear properties are to be measured, is placed on the fused quartz surface, this causes a change in the amplitude and phase of the reflected wave. By using the balance method shown by Fig. 7, in which two identical fused quartz rods are used, one of which has a liquid layer and the other does not, and by using a phase shifting network and an attenuator to balance out

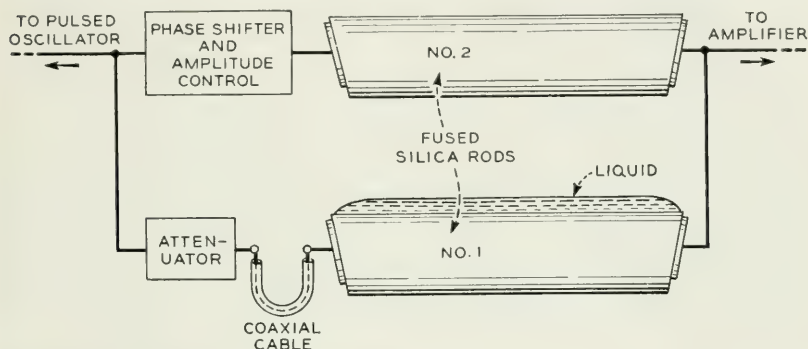


Fig. 7—Method for obtaining resistance and reactance terms for high frequency shear reflection method.

pulses, the shear impedance of the liquid can be determined. If R is the loss per reflection expressed as a current ratio, θ the change in phase angle required to rebalance the circuit and φ the angle between the wave normal and the reflecting surface, it can be shown⁸ that the shear impedance of the liquid is

$$Z_M = R_M + jX_M = Z_q \cos \varphi \left[\frac{1 - R^2 + 2jR \sin \theta}{1 + R^2 + 2R \cos \theta} \right] \quad (7)$$

where Z_q is the impedance ρv for shear waves in the quartz. This is equal to

$$Z_q = 2.20 \times 3.76 \times 10^5 = 8.27 \times 10^5 \text{ mechanical ohms}$$

Since this impedance is much larger than that of the liquids that are to be measured, the sensitivity is increased by making φ large. In practice φ was taken as 80° . This method is applicable from 3 mc up to 100 mc

and complements the other methods. Fig. 8 shows a photograph of the equipment.

III. MEASUREMENTS OF POLYMERS IN SOLUTION

When such methods are applied to a polymer solution, it is found that the resistance and reactance components are no longer equal but the resistance is invariably larger than the reactance. This indicates the presence of a shear elasticity in the solution. If the molecules have a single relaxation frequency, it has been found that the shear properties of the liquid can be represented by a stress-strain equation of the type

$$T = \eta_A \frac{\partial S}{\partial t} + \frac{1}{\frac{1}{\eta_B \frac{\partial S}{\partial t}} + \frac{1}{\mu_B S}} \quad (8)$$

where T is the shearing stress, S the shearing strain, η_A the solvent viscosity, η_B a molecular viscosity of some particular motion of the chain which disappears when the reactance of the chain stiffness μ_B of this motion is low enough so that the motion can follow the applied shearing stress at the frequency of the measurement. When this type of mechanism is present in the liquid, it has been shown⁹ that the impedance the liquid presents to the crystals is

$$Z_0 = R_M + jX_M = \frac{\rho\mu_B\eta_B^2 + j\left[\omega\rho\eta_A\eta_B^2 + \frac{\rho\mu_B^2}{\omega}(\eta_A + \eta_B)\right]}{\eta_B^2 + \mu_B^2/\omega} \quad (9)$$

Fig. 9 shows a plot of the resistance and reactance components of an assumed solution having a single relaxation frequency, and a viscosity 30 times the solvent viscosity. At very low frequencies, the resistance and reactance follow that of a solution, but for frequencies comparable with the relaxation frequency, the resistance becomes larger than the reactance while for very high frequencies the two come together on a line determined by the solvent viscosity. If there is more than one relaxation frequency, the resistance and reactance may coalesce for several intermediate stages. A continuous distribution would give a definite relation between the frequency dependence of resistance and reactance.

The torsional crystal and the shear wave reflection method have been applied to long chains of polyisobutylene dissolved in various sol-

⁹ W. P. Mason, *Piezoelectric Crystals and Their Application to Ultrasonics*, D. van Nostrand, 1950, p. 353.

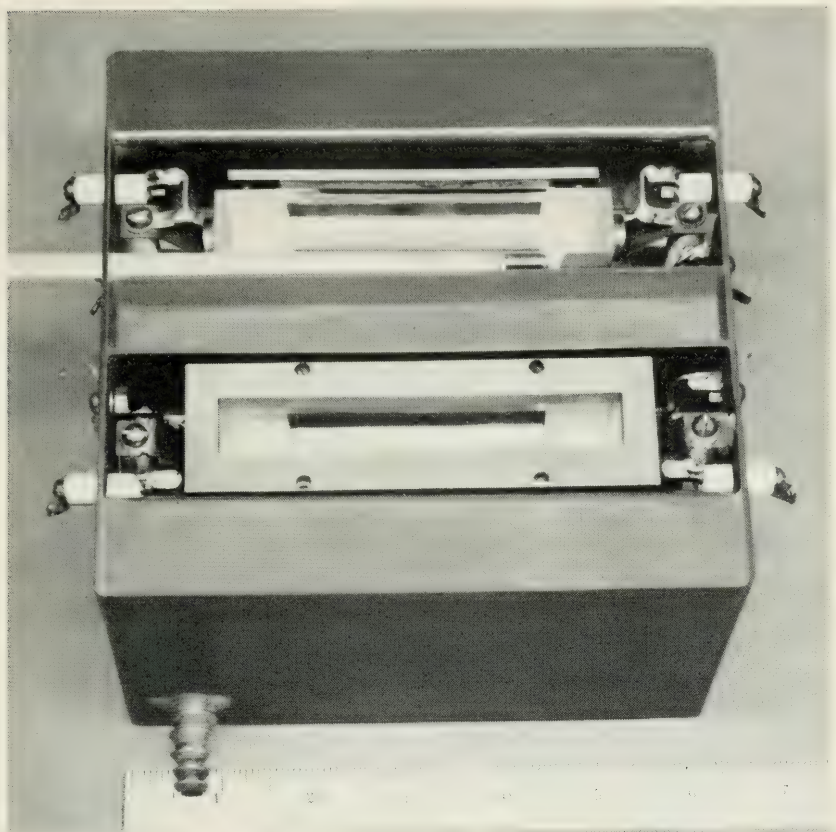
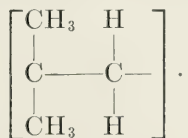


Fig. 8—Dual reflecting block assembly for measuring high frequency shear impedances of liquids.

vents with concentrations ranging from zero per cent to 10 per cent. Polyisobutylene is a polymer molecule having the chemical formula



Non-planar zigzag segments can be expected in the liquid state. Fig. 10 shows measured curves for 20 kc of the resistance and reactance for solutions of viscosity average molecular weight of 3,930,000 in cyclohexane. Four values of concentration were used and two temperatures were measured. For pure cyclohexane, the resistance and reactance

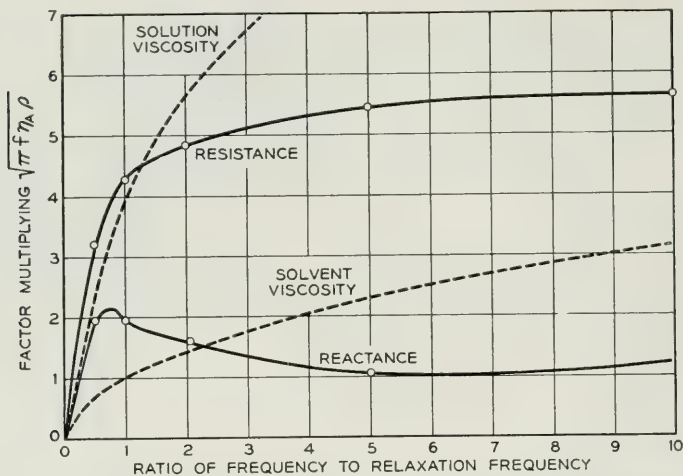


Fig. 9—Resistance and reactance components of a solution having a single relaxation frequency and a solution viscosity 30 times the solvent viscosity.

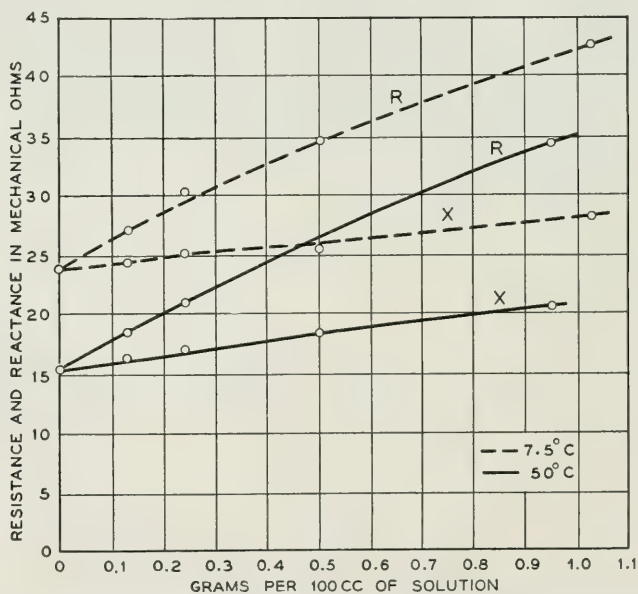


Fig. 10—Shear resistance and reactance components of a solution of polyisobutylene (molecular weight of 3,930,000) in cyclohexane plotted as a function of grams per cc of solution.

components are equal but as the percentage of polyisobutylene is increased, the resistance increases more rapidly than the reactance.

By solving equation (9) for η_A , η_B and μ_B in terms of R and X measured at one frequency and $\eta_A + \eta_B$ the solution viscosity, we find

$$\eta_A = \frac{2RX}{\omega\rho} - \frac{(R^2 - X^2)/\omega\rho}{\omega\rho(\eta_A + \eta_B) - 2RX}; \quad \eta_B = (\eta_A + \eta_B) - \eta_A \quad (10)$$

$$\mu_B = \frac{(R^2 - X^2)\omega\eta_B}{\omega\rho(\eta_A + \eta_B) - 2RX}$$

Applying these formulae to the measured results, the curves of Fig. 11 result. The shear elasticity is directly proportional to the concentration, the viscosity η_A is only slightly larger than the solvent viscosity while the main part of the measured viscosity resides in η_B the viscosity associated with chain motion. Fig. 12 shows these three quantities for a one per cent solution measured as a function of temperature. The apparent

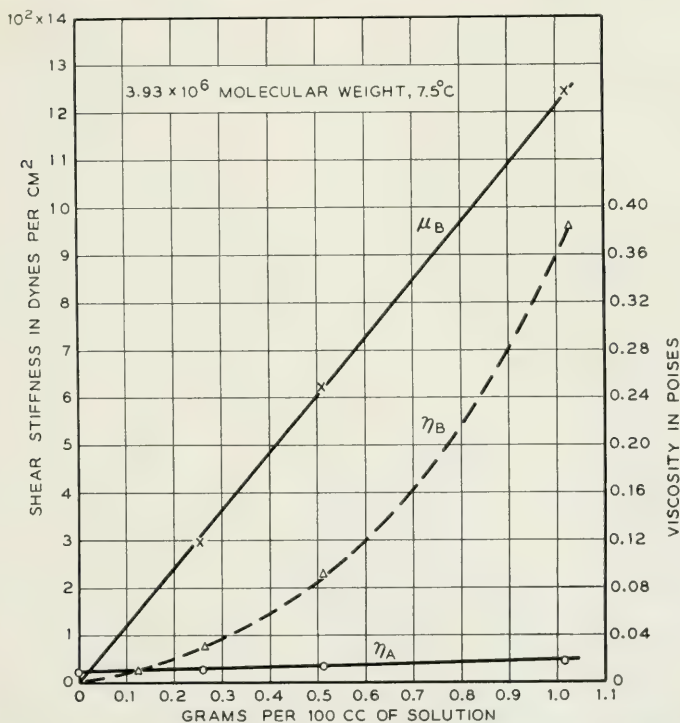


Fig. 11—Shear stiffness, series viscosity η_A and molecular viscosity η_B for polyisobutylene (molecular weight of 3,930,000) in cyclohexane plotted as a function of grams per 100 cc of solution.

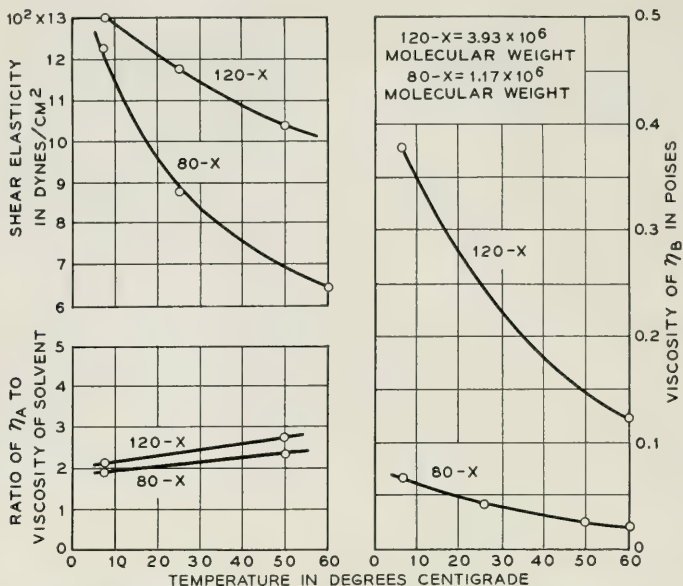


Fig. 12—Shear stiffness, series viscosity and molecular viscosity plotted as a function of temperature for two molecular weight solutions of polyisobutylene in cyclohexane.

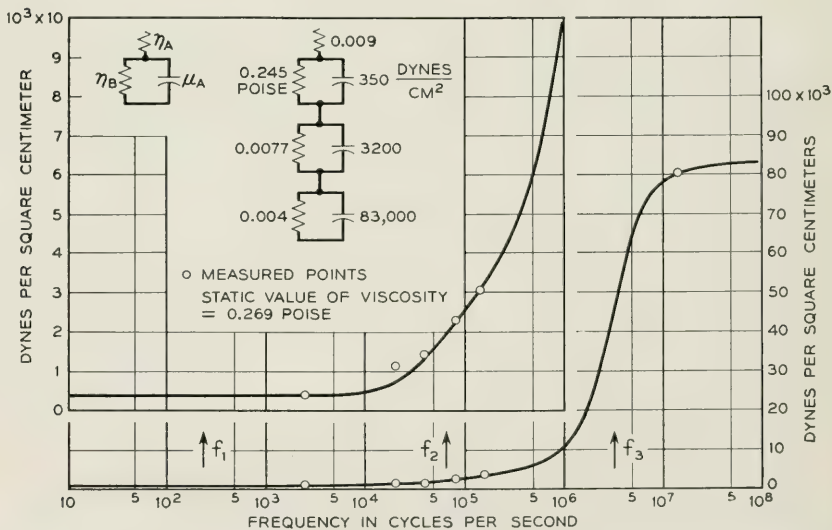


Fig. 13—Shear elasticity for a one per cent solution of polyisobutylene as a function of frequency for 25°C.

stiffness decreases with increase in temperature for a single frequency measurement. However, when measurements were made at 20, 40, 80 and 150 kc it was found that the elasticity was a function of the frequency, which indicates the presence of more than one relaxation and complicates the determination of the temperature relationships. Fig. 13 shows measurements of the shear elasticity over a frequency range from 2.5 kc¹⁰ up to 14 megacycles for 25°C. There is a gradual rise up to about 300 kilocycles after which there is a sharp break to a stiffness of about 90,000 dynes/cm² for a 1 per cent solution of the 3,930,000 molecular weight polymer in cyclohexane. If one analyzes the frequency variation of the elasticity he finds that it can be fitted by three relaxation frequencies, one having a frequency of 230 cycles, one around 66,000 cycles and one around 4 megacycles. A possibility exists for a fourth relaxation. The lowest relaxation is thought to be a configurational relaxation of all the elements of the chain. The highest one appears to be a relaxation of the twisting motion of the smallest segment of the chain such as a Kuhn segment. The intermediate relaxation appears to be due to the motion of the ends of the smallest chain segment from one position of entanglement to an adjacent position. This interpretation is based partly on the fact that the associated viscosity of the motion is very similar to that for the relaxation of the twisting motion of the smallest chain segment and partly from data presented in the next section on pure polymer liquids which shows a lower frequency relaxation agreeing in frequency asymptotically with this one, which involves chain motions of approximately 30 to 40 chain elements. Temperature variations of these elastic components show that the lowest relaxation mechanism has a stiffness that increases slightly with temperature in agreement with the kinetic theory of elasticity. The corresponding viscosity (η_2) which comprises most of the viscosity for a solution, when plotted against the reciprocal of the temperature, as shown by Fig. 14, indicates an activation energy of 3.9 kilocalories per mole which is slightly higher than that of the solvent cyclohexane alone, which is about 3.2 kilocalories per mole. This difference of 0.7 kilocalories presumably represents the added energy required to bend the chain in its configurational motion. Measurements with another chain length of 1.18×10^6 molecular weight showed that the stiffness of the lowest (configurational) relaxation decreased from 310 to 160 indicating that the stiffness of this motion is approximately proportional to the square root

¹⁰ The lowest frequency, 2.5 kc, was measured by means of a quartz crystal tuning fork which will be described in another paper. This instrument makes possible the direct measurement of configurational elasticities.

of the molecular weight. The viscosity decreased by a factor of 6.25 and consequently the relaxation frequency increases from 230 to 660 cycles.

The second "entanglement" relaxation has a stiffness of about 3100 dynes/cm² for a 1 gram per cc solution of the 3.93×10^6 molecular weight solution and about 2650 dynes/cm² for the 1.18×10^6 molecular weight solution. The variation with temperature, if any, is small. The corresponding viscosities η_3 for the two solutions are nearly equal as shown by Fig. 14, and have an activation energy of 4.25 kilocalories per mole. The final high frequency "short segment" relaxation has a high stiffness of 83,000 dynes/cm² for a 1 per cent solution of 3.93×10^6 molecular weight. The corresponding viscosities for the two solutions shown by Fig. 14 have nearly identical values and an activation energy of 4.25 kilocalories per mole, i.e. very closely equal to the "entanglement" relaxation viscosity.

These upper two relaxations persist in pure liquid polymers as discussed in the next section, although they are spread out over a small range of relaxation time values. The highest one can be traced in measurements of mechanical properties of solid plastics such as polyethylene and nylon which indicates that these materials should have rubber like

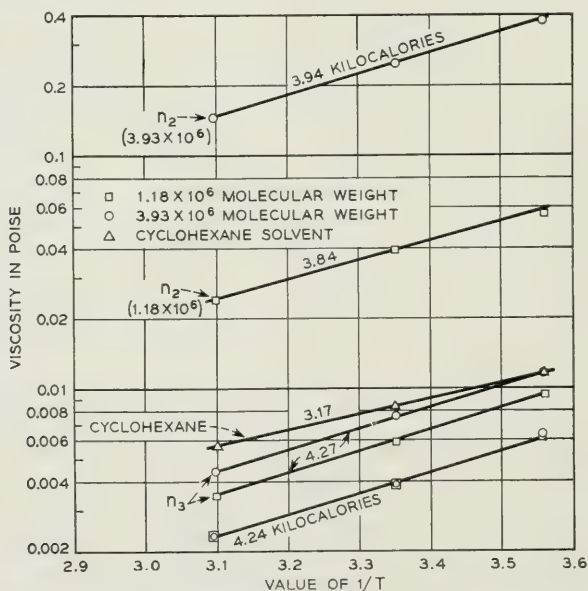


Fig. 14—Viscosities for the three components of the motion and for the solvent plotted against the inverse of the absolute temperature.

TABLE I

Liquid	Average Molecular Weight	Density				Melt Viscosity Poise				No. of Chain Segments
		15°C	25°C	35°C	50°C	15°C	25°C	35°C	50°C	
A	904	0.83	0.826	0.821	0.814	1.10	0.67	0.33	0.16	16.2
B	1,220	0.846	0.842	0.837	0.83	2.64	1.40	0.65	0.297	22.2
C	1,590	0.856	0.849	0.846	0.840	8.4	4.0	2.1	0.87	28.4
D	2,450	0.872	0.866	0.863	0.857	100	38	15.9	4.9	44.0
E	3,520	0.881	0.877	0.872	0.866	580	216	95	30	63.0
F	4,550	0.89	0.886	0.882	0.875	1,960	737	320	98	81.4
G	5,590	0.897	0.892	0.887	0.882	4,600	1,840	740	220	100
H	10,380	0.912	0.908	0.904	0.896	11,500	4,600	2,050	625	186

Measurement at 24 megacycles $\frac{\mu \text{ dynes}}{\text{cm}^2}$; η in poise

Liquid	15°C		25°C		35°C		50°C	
	μ	η	μ	η	μ	η	μ	η
A	0.482×10^9	1.07	0.4×10^9	0.68	0.35×10^9	0.39	0.1×10^9	0.13
B	1.08×10^9	2.64	0.8×10^9	1.39	0.58×10^9	0.7	0.17×10^9	0.293
C	1.65×10^9	5.12	1.2×10^9	2.57	0.885×10^9	1.99	0.27×10^9	0.881
D	2.71×10^9	15.7	2.0×10^9	8.1	1.67×10^9	4.15	0.98×10^9	1.77
E	3.81×10^9	40	3.0×10^9	20	2.17×10^9	10.6	1.2×10^9	4.62
F	5.48×10^9	73	4.22×10^9	38.3	2.8×10^9	20.2	1.6×10^9	8.8
G	5.9×10^9	107	4.78×10^9	51	3.74×10^9	29.5	2.5×10^9	13.2
H	6.9×10^9	119	5.55×10^9	65.8	4.22×10^9	31.5	2.9×10^9	14.8

Measurements at 14 megacycles

A	0.38×10^9	0.975	0.3×10^9	0.6	0.22×10^9	0.35	0.12×10^9	0.15
B	0.475×10^9	2.53	0.35×10^9	1.4	0.25×10^9	0.74	0.18×10^9	0.53
C	0.68×10^9	3.86	0.61×10^9	3.4	0.3×10^9	1.62	0.25×10^9	0.56
D	2.32×10^9	16.3	1.7×10^9	10	0.94×10^9	4.75	0.39×10^9	1.86
E	3.8×10^9	56.2	2.8×10^9	24.25	2.0×10^9	12.3	0.855×10^9	5.7
F	4.75×10^9	90.4	3.6×10^9	48.3	2.65×10^9	26.6	1.47×10^9	10.8
G	6.03×10^9	136.5	4.6×10^9	81.4	3.24×10^9	39.6	2.0×10^9	15
H	6.64×10^9	160	5.3×10^9	93.5	3.98×10^9	54.2	2.3×10^9	18.5

Measurement at 4.5 megacycles

A	0.34×10^9	1.18	0.19×10^9	0.64	0.18×10^9	0.34	0.09×10^9	0.17
B	0.55×10^9	2.65	0.21×10^9	1.33	0.2×10^9	0.68	0.1×10^9	0.20
C								
D	1.57×10^9	24	0.96×10^9	11.7	0.72×10^9	5.52	0.34×10^9	2.1
E	2.79×10^9	76	1.9×10^9	37.4	1.43×10^9	16.5	0.9×10^9	5.15
F	3.57×10^9	124	2.5×10^9	61.2	1.86×10^9	31	1.04×10^9	12.8
G	4.4×10^9	176	3.4×10^9	98.5	2.6×10^9	54	1.6×10^9	22.6
H	5.65×10^9	186	4.3×10^9	124	2.8×10^9	76	1.8×10^9	28.8

response to mechanical shocks of very short duration. The lowest frequency configurational relaxation is spread over a wide spectrum of relaxation times in pure liquids.

Measurements¹¹ of these and other chains in various solvents have also been made and the results are discussed, from a chemical point of view, in a companion paper by W. O. Baker and J. H. Heiss. It is shown that the stiffnesses vary with the polymer chain and the solvent used.

IV. MEASUREMENTS OF PURE LIQUID POLYMERS

A. Shear Wave Measurements in Liquid Polymers

Similar shear wave measurements have been made for pure polyisobutylene liquids of molecular weights from 904 to 10,380 (i.e. from 16 chain elements to 186 chain elements), by the techniques described in Section II. Some of these results have been discussed in reference (8) but the much more comprehensive measurements made since require some revisions of the original conclusions.

The easiest data to interpret are the high-frequency data obtained by the shear wave reflectance method. The data of Table I give measurements of 8 liquids varying in average molecular weight from 900 to 10,380, at three frequencies and four temperatures. If we plot for example the Maxwell shear stiffness and viscosity for the three frequencies and for 25°C as a function of the number of chain elements (here a chain element is taken as two adjacent carbon atoms one of which has two methyl groups attached and the other two hydrogens) the 4.5-mc measurements are shown by the triangles of Fig. 15. The 14 megacycle measurements are shown by the circles and the 24-mc measurements by the squares.

An attempt was made to fit these measurements with a two relaxation mechanism shown by the figure with two stiffnesses which are taken to be independent of the molecular weight and equal respectively to 1.2×10^8 dynes/cm² and 6×10^9 dynes/cm². The best fit is obtained by taking the two viscosities η_1 and η_2 equal and these are adjusted for the different molecular weights in such a manner as to best fit the experimental curve. A fair agreement is obtained except for the range from 60 to 90 chain elements where the two relaxation model gives too rapid an increase of stiffness with increase in the number of chain elements and at the high molecular weight viscosity range where the viscosity shows a dispersion in values but the model does not. The sum of the two vis-

¹¹ These results on the mechanical impedance of long chain molecules in solvents have been presented at the XIIth International Congress of Pure and Applied Chemistry by W. O. Baker, W. P. Mason and J. H. Heiss, Sept. 13, 1951.

cosities η_1 and η_2 assumed as a function of molecular weight is shown by Fig. 16. The log of the viscosity starts proportional to the molecular weight but above a molecular weight of 2,400 the increase is very slow and becomes asymptotic to a value of 240 poises. An equation which fits the increase in viscosity with molecular weight is

$$\eta_D = K e^{11.8 \tanh Z/2370}$$

where Z is the molecular weight. The solid line shows a plot of this curve and the circles are the assumed values to obtain a best fit to the meas-

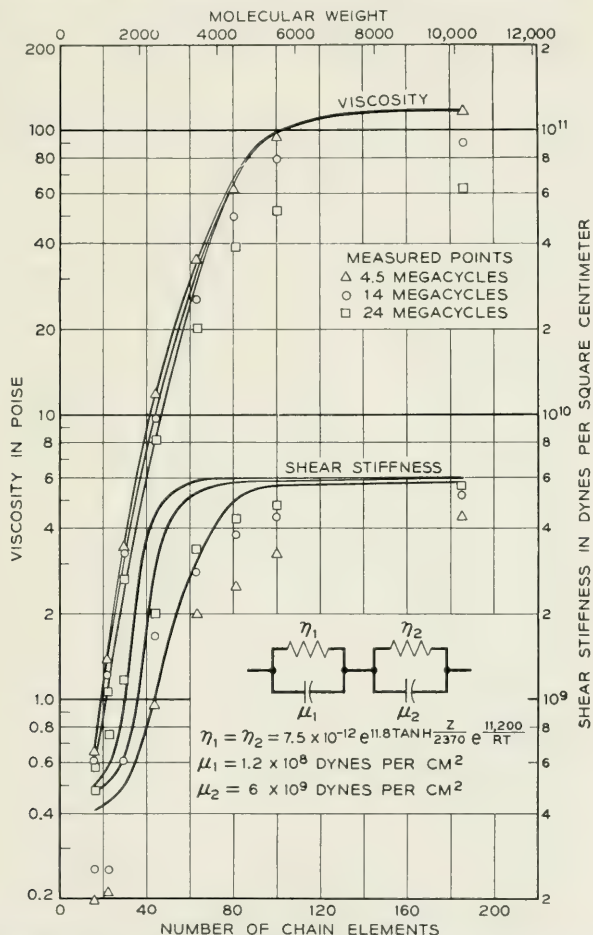


Fig. 15—Measured values of high frequency shear viscosity and elasticity for 25°C and three frequencies plotted against molecular weight. Solid lines are best fit obtained by a two relaxation mechanism having the element values shown by the figure.

ured values. This equation indicates that when $Z = 2,370$ or 43 chain elements the viscosity increases only a small amount more by a chain articulation effect and hence in this high frequency range we are dealing with a chain length of about 40 elements or 80 carbon atoms. This is checked also by a comparison of the static and dynamic viscosity. The total dynamic viscosity due to the two relaxation mechanisms compared to the static viscosity does not differ markedly until the number of chain elements is more than 40. Above this value other motions than that of the shortest chain segment can take place and can add to the dynamic viscosity. The static viscosity fits an equation of the same sort, but the indicated chain length for the viscous motion is about $\frac{5}{3}$ times that of the shortest segment.

When a similar process is carried out over the temperature range the equations of Fig. 16 are obtained. The static viscosity has an activation energy of 16 kilocalories per mole, while the dynamic viscosity has an activation energy of about 11.2 kilocalories per mole.

The relaxation frequencies for the two components are plotted as a function of the number of chain segments by the solid lines of Fig. 17.

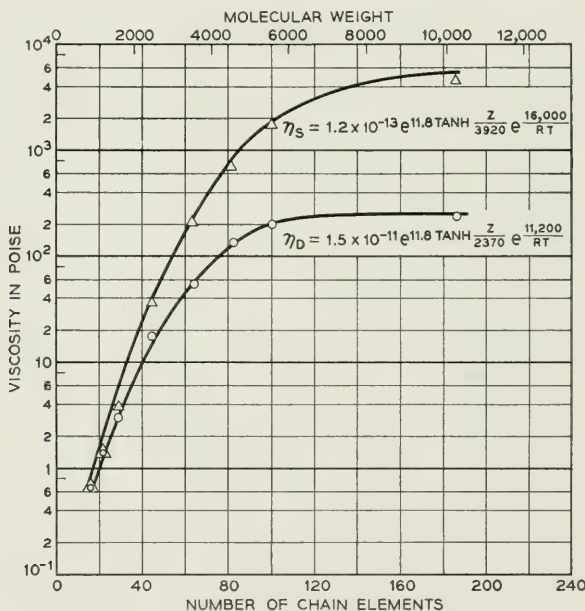


Fig. 16—Triangles are measured static viscosities and circles are dynamic viscosities plotted as a function of molecular weight. Solid lines are a plot of the equations given for static and dynamic viscosities.

For long chain segments the values become asymptotic to 8×10^6 cycles and 160,000 cycles which are not far from the two highest relaxation frequencies obtained from the solution measurements of Section III. Hence it appears likely that these relaxations are due to the "entanglement" motion and the twisting motion of the shortest chain segment. The increased activation energy is due to the fact that more energy has to be applied to the chain segment to break it loose from its equilibrium position when it is surrounded by adjacent polyisobutylene molecules than when it is surrounded by cyclohexane molecules. The stiffness of the chain is due more to the slope of the potential well than to any intrinsic chain stiffness as is shown by Fig. 18, which shows the two stiffnesses as a function of temperature. These values are obtained by fitting the 15°, 35° and 50° data in a similar manner to that used for the 25°

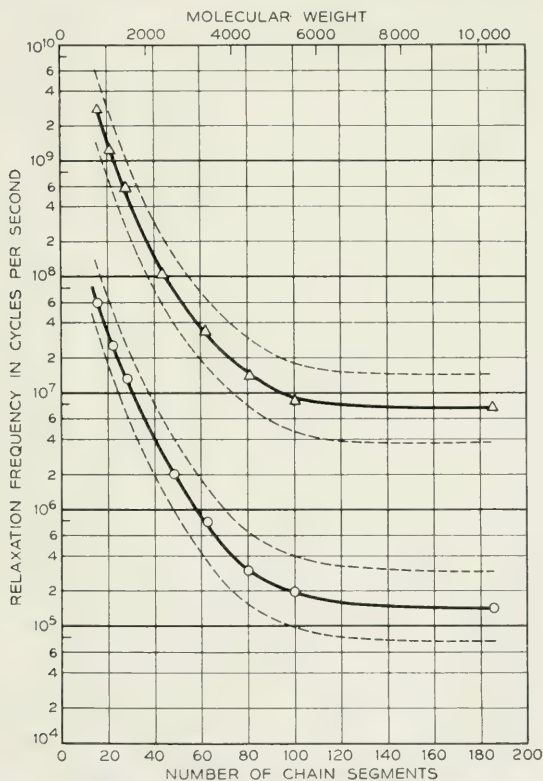


Fig. 17—Solid lines are mean values of relaxation frequencies for the two mechanisms plotted against molecular weight. Dotted lines indicate limits of regions assumed to obtain a better fit to the measured values.

data, and the activation energy of 11.2 kilocalories per mole is obtained in a similar manner.

Due to the closeness of the surrounding polyisobutylene molecules, one would expect that the relaxation frequencies would not have discrete values but would be spread about the center value in some sort of a Gaussian distribution. If we approximate this by representing each region by two relaxations, one-half, and the other twice the frequency of the mean value, as shown by the dotted lines of Fig. 17, the agreement with the measured values of Fig. 15 is considerably better as shown by Fig. 19. A wider distribution yet is indicated.

For molecular weights greater than 2,000, the shortest chain segment viscosity begins to diverge from the static viscosity indicating that there are other relaxations for these longer chains. Some data for the three longest chain polymers, F, G and H have been obtained by the torsional rod method and the results are given in Table II. These data are plotted on Fig. 20 as a ratio of dynamic to static viscosity plotted

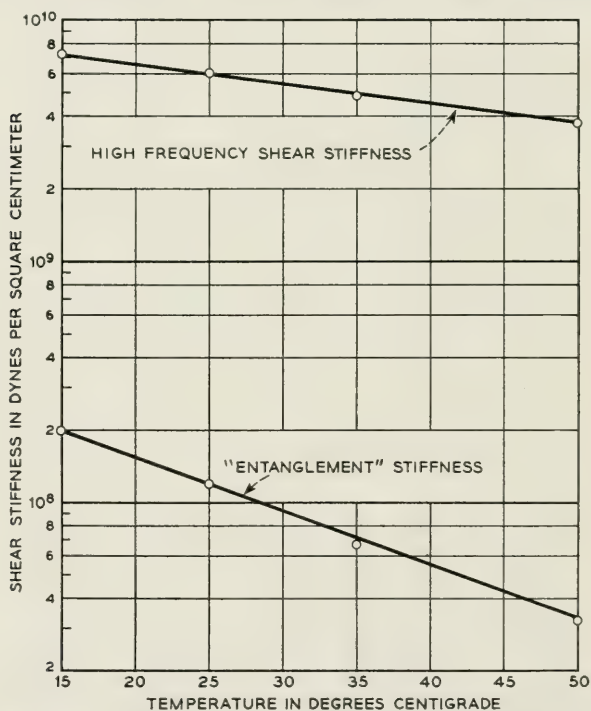


Fig. 18—Variations of high frequency shear stiffness and "entanglement" stiffness plotted as a function of the temperature.

against the frequency times the static viscosity. All the viscosity data can be represented within the experimental error by a single curve, but the stiffness curves appear to require different curves for different temperatures. On analyzing the data in terms of a distribution of relaxation frequencies, a single curve for all temperatures could be obtained if the stiffness of each mechanism were independent of the temperature and the relaxation frequency were inversely proportional to the static viscosity, i.e., had an activation energy variation equal to that for the static viscosity for each relaxation mechanism. This condition holds

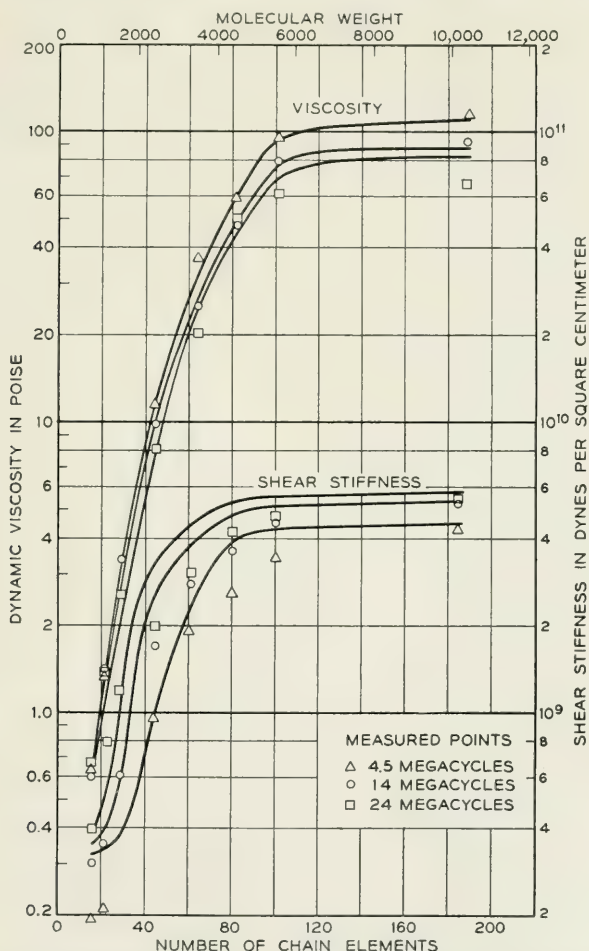


Fig. 19—Curves showing better fit to experimental data obtained by assuming the relaxation regions shown by Fig. 17.

TABLE II
Measurement for Polymer F—Molecular weight = 4550

Frequency Kilocycles	15°C		25°C		35°C		45°C		55°C		65°C	
	μ	η	μ	η	μ	η	μ	η	μ	η	μ	η
25	2.9×10^8	600	1.45×10^8	270	0.78×10^8	145	0.44×10^8	77	0.26×10^8	43	0.17×10^8	24
40	3.8×10^8	550	1.7×10^8	250	0.86×10^8	125	0.47×10^8	66	0.28×10^8	38	0.20×10^8	22
52	4.0×10^8	470	1.9×10^8	220	0.78×10^8	110	0.54×10^8	59	0.33×10^8	34	0.21×10^8	20
140	5.8×10^8	380	3.0×10^8	180	1.5×10^8	91	0.71×10^8	48	0.44×10^8	26	0.28×10^8	17

Polymer G—Molecular weight = 5590

Frequency Kilocycles	24.5°C		65°C	
	μ	η	μ	η
27.5	2.0×10^8	510	1.9×10^7	38.4
40.96	2.95×10^8	480		
54.23	3.60×10^8	456	2.93×10^7	36.1

Polymer H—Molecular weight = 10,380

Frequency Kilocycles	55°C		67°C	
	μ	η	μ	η
32.2	3.9×10^7	150	2.9×10^7	94
40.5	4.8×10^7	137	3.5×10^7	86
47.0	5.3×10^7	126	3.9×10^7	79

quite well for frequencies much lower than the relaxation frequencies of the smallest chain segment, but as the frequency approaches these relaxation frequencies, the stiffness of these polymers increases as the temperature decreases.

A fair approximation to these measured values is obtained by assuming one more "configurational" relaxation frequency in addition to the two smallest segment relaxations discussed previously. Fig. 21 shows calculations of the ratio of dynamic to static viscosity and the shear stiffness for 65°C and 25°C. The lowest relaxation frequency is assumed to have a stiffness of 6.3×10^6 dynes/cm² and a viscosity of 20 poises at 65°C. For 25°C the stiffness of 2×10^7 dynes/cm² is assumed and an activation energy of 17.3 kilocalories gives the component a viscosity of 607 poises at 25°C, and a relaxation frequency of 5,250 cycles. The average value of 16 kilocalories for the static viscosity is a result of the sum of the variation due to the two components. Although the agreement can be improved by assuming distributions of relaxation frequencies centered around these three primary frequencies, there does not seem to be much doubt of the existence of these primary relaxation

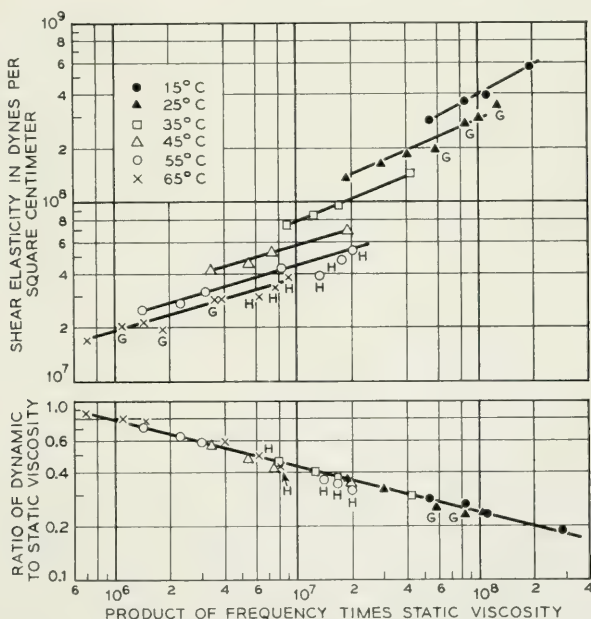


Fig. 20—Plot of ratio of dynamic to static viscosity and the corresponding intermediate frequency shear stiffnesses as a function of temperature and product of frequency times static viscosity.

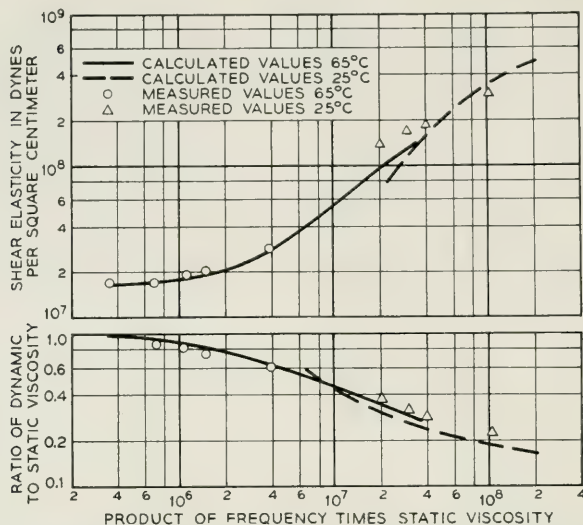
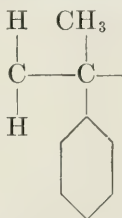


Fig. 21—Calculated values of stiffness and viscosity obtained by adding a single "configurational" relaxation frequency to the two short chain relaxations obtained from high frequency measurements.

mechanisms which show up as discrete relaxations in long chain molecules in solution.

Some measurements have also been made to determine the effect of chemical substitutions in the polymer chains. In a previous paper,* the high frequency properties of poly- α -methyl styrene were discussed. This material has the polyisobutylene chain but with one methyl replaced by a phenyl so that its chain becomes

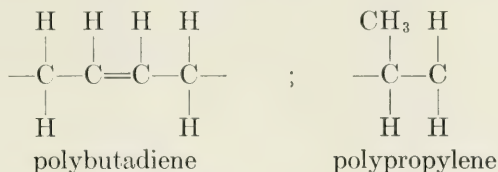


For low molecular weights this material is liquid. The shear stiffness of this liquid is somewhat higher than for polyisobutylene but has about the same change with temperature. The variation of the high-frequency viscosity, however, is much larger for poly- α -methyl styrene than for polyisobutylene, and corresponds to an activation energy of 23.6 kilocalories. The relaxation region for the shortest chain motion is much

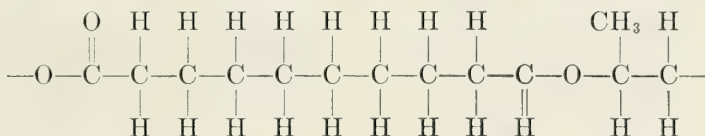
* See footnote on page 132.

narrower than in polyisobutylene, which correlates with the smaller steric hindrance.

Measurements have also been made in the 70-ke to 140-ke range for two non polar liquids, polybutadiene and polypropylene and one polar liquid, polypropylene sebacate. The first two have the formulae



while the polar liquid, polypropylene sebacate, has the formulae



The measured results are given by Table III. These data are plotted on Fig. 22 as a function of the product of frequency times static viscosity. For comparison the data for polyisobutylene polymer F is also plotted. By extrapolating to low values of the product-frequency times static viscosity it is seen that the low frequency "quasi-configurational" stiffnesses of these liquids run from 10^6 to 1.5×10^7 dynes per square centimeter. Polyisobutylene has the greatest stiffness for any value of frequency times viscosity while polybutadiene has the least. This reflects the greatest steric hindrance of polyisobutylene and the smallest for polybutadiene which has only hydrogens connected to the carbon chain atoms. Another consequence of the larger steric hindrance of the CH_3 groups of polyisobutylene is that the viscosity associated with the shortest chain segment motion is largest for polyisobutylene and smallest for polybutadiene as can be seen from the ratio of dynamic to static viscosities for high values of frequency times static viscosity.

The activation energy for static viscosity are for polypropylene, polypropylene sebacate and polybutadiene respectively 21.2, 12 and 8 kilocalories compared to 16 kilocalories per mole for polyisobutylene. The high-frequency activation energies as determined by the dynamic measurements are respectively 11.8, 4.6 and 1.4 kilocalories for polypropylene, polypropylene sebacate and polybutadiene. The differences between the activation energy for static flow and that for dynamic flow are respectively 9.6, 7.4 and 6.6 kilocalories, which values are all higher than

TABLE III

Temp. °C	Density ρ	76.5 Kc Measurements				142.6 Kc Measurements				
		Poise Static Viscosity η_s	Maxwell		η_D/η_s	Maxwell		η_D/η_s	$f\eta_s$	
			η_D poise	μ dynes/cm ²		η_D poise	μ dynes/cm ²			
Polybutadiene										
10	0.877	740	25.5	7.83 $\times 10^6$	0.0355	5.66 $\times 10^7$	14.65	9.96 $\times 10^6$	0.0196	10.55 $\times 10^7$
20	0.873	450	21.1	6.38 $\times 10^6$	0.0469	3.44 $\times 10^7$	13.45	7.91 $\times 10^6$	0.0299	6.41 $\times 10^7$
30	0.869	288	17.45	5.3 $\times 10^6$	0.0605	2.21 $\times 10^7$	12.4	6.67 $\times 10^6$	0.043	4.11 $\times 10^7$
40	0.865	189	15.3	4.67 $\times 10^6$	0.081	1.44 $\times 10^7$	11.05	5.96 $\times 10^6$	0.0583	2.69 $\times 10^7$
50	0.861	128	11.9	3.92 $\times 10^6$	0.093	0.98 $\times 10^7$	9.16	5.21 $\times 10^6$	0.0715	1.82 $\times 10^7$
60	0.857	88	9.85	3.55 $\times 10^6$	0.112	0.67 $\times 10^7$	7.51	4.54 $\times 10^6$	0.0855	1.25 $\times 10^7$
Polypropylene										
76.5 Kc Measurements										
45	0.849	30,000	380	3.56 $\times 10^8$	0.01265	2.3 $\times 10^9$	354.5	6.69 $\times 10^8$	0.0118	4.26 $\times 10^9$
55	0.842	10,900	160.2	1.4 $\times 10^8$	0.0153	0.832 $\times 10^9$	167	3.18 $\times 10^8$	0.0153	1.55 $\times 10^9$
65	0.836	4,200	89.8	0.786 $\times 10^8$	0.0214	0.321 $\times 10^9$	79.1	1.46 $\times 10^8$	0.0188	0.6 $\times 10^9$
75	0.830	1,600	50.8	0.444 $\times 10^8$	0.0318	0.128 $\times 10^9$	46.6	0.998 $\times 10^8$	0.0292	0.23 $\times 10^9$
85	0.824	700	31.2	0.264 $\times 10^8$	0.0446	0.053 $\times 10^9$	28.4	0.51 $\times 10^8$	0.0405	0.10 $\times 10^9$
Polypropylene Sebacate										
77.1 Kc Measurements										
5	1.076	9,200	95.2	4.88 $\times 10^7$	0.0103	7.1 $\times 10^8$	74.6	7.45 $\times 10^7$	0.0081	1.3 $\times 10^9$
15	1.068	4,200	67.6	2.94 $\times 10^7$	0.0161	3.24 $\times 10^8$	54.9	4.28 $\times 10^7$	0.0131	0.595 $\times 10^9$
25	1.060	2,060	51.4	2.01 $\times 10^7$	0.0249	1.6 $\times 10^8$	41.3	2.82 $\times 10^7$	0.02	0.292 $\times 10^9$
35	1.051	1,050	43.1	1.54 $\times 10^7$	0.0411	0.81 $\times 10^8$	33.2	1.97 $\times 10^7$	0.0316	0.148 $\times 10^9$
45	1.042	580	38.8	1.24 $\times 10^7$	0.0669	0.45 $\times 10^8$	26.9	1.69 $\times 10^7$	0.0464	0.082 $\times 10^9$
55	1.034	320	32.8	1.01 $\times 10^7$	0.1023	0.25 $\times 10^8$	24.2	1.34 $\times 10^7$	0.0756	0.045 $\times 10^9$
65	1.026	183	26.8	0.88 $\times 10^7$	0.147	0.14 $\times 10^8$	22.9	1.25 $\times 10^7$	0.125	0.026 $\times 10^9$

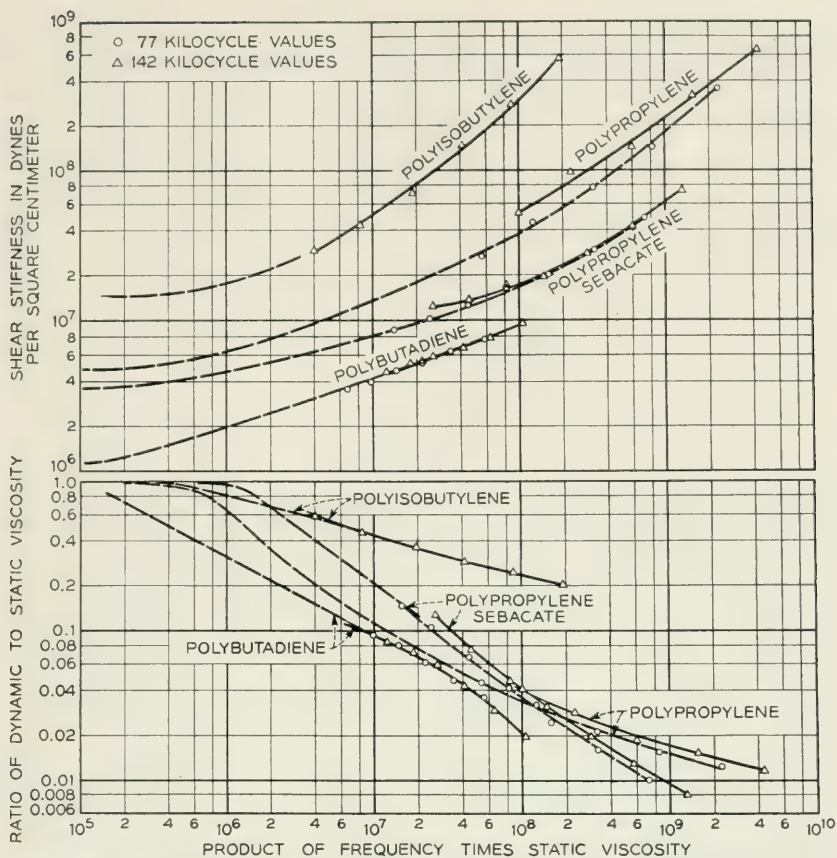


Fig. 22—Ratio of dynamic to static viscosity and the shear stiffness for four polymer liquids plotted against product of frequency and static viscosity.

the 4.8 kilocalories for polyisobutylene. This presumably indicates that there is more of a difference between the viscosity flow segment and the shortest chain segment in these materials than in polyisobutylene. Since no measurements are available over a range of molecular weights, no direct evidence has been obtained for the various chain lengths.

B. Longitudinal Wave Measurements in Liquid Polymers

Since the increase in shear elasticity for the highest relaxation frequency is so large, it should also appear in longitudinal wave measurements. Fig. 23 shows a calculation for the 5590 molecular weight liquid of the longitudinal velocity assuming that the Lamé λ elastic constant is independent of frequency and that all the variation occurs in the shear

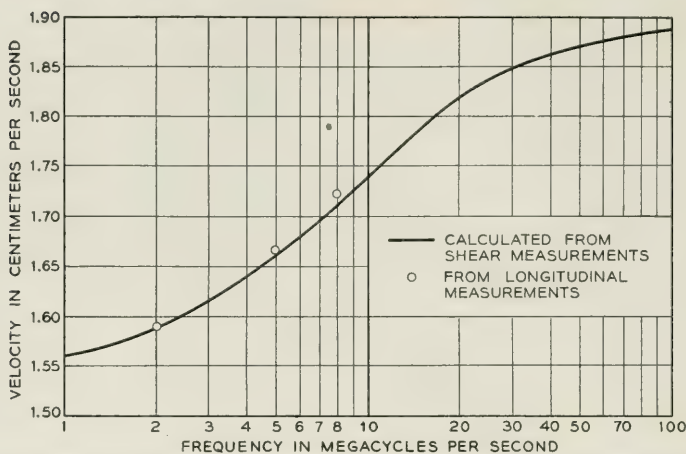


Fig. 23—Relation between measured longitudinal velocity for polyisobutylene of molecular weight 5590 and that calculated from shear stiffness measurements assuming the Lamé λ elastic constant is independent of frequency.

constant as determined by the shear measurements. The points are velocities measured for longitudinal waves and as can be seen, the measurements agree closely with the calculated values. A slightly better agreement would be obtained if λ increased by a small amount as the frequency increased. As discussed in the next section there is some experimental evidence for an increase in λ in nylon 6-6 and in polyethylene.

The question also arises as to how much of the attenuation is due to shear mechanisms and how much due to pure compressional effects. From longitudinal velocity and attenuation measurements at 30°C for the polymers E, F and G of Table I, the values of $\lambda + 2\mu$ and $\chi + 2\eta$ can be determined and are shown by Table IV. The values of μ and η can be obtained from Table I by interpolation and are given in columns

TABLE IV

Polymer	$\lambda + 2\mu$ dynes/cm ²	$\chi + 2\eta$ poise	μ dynes/cm ²	η poise	λ dynes/cm ²	χ
5 megacycles						
E	1.92×10^{10}	62	0.17×10^{10}	26	1.60×10^{10}	10
F	2.26	107	0.23×10^{10}	46	1.70×10^{10}	15
G	2.38	170	0.31×10^{10}	75	1.76×10^{10}	20
8 megacycles						
E	2.01×10^{10}	50	0.20×10^{10}	22	1.61×10^{10}	6
F	2.26	93	0.27×10^{10}	41	1.72×10^{10}	11
G	2.56	155	0.34×10^{10}	69	1.88×10^{10}	17

4 and 5. Columns 6 and 7 show the values of λ and χ , the compressional components. A definite longitudinal compressional viscosity is indicated which however is somewhat smaller than the shear viscosity η .

V. MEASUREMENTS FOR SOLID POLYMERS

Recently two new methods have been devised for accurately measuring the properties of solid plastics in the ultrasonic frequency range

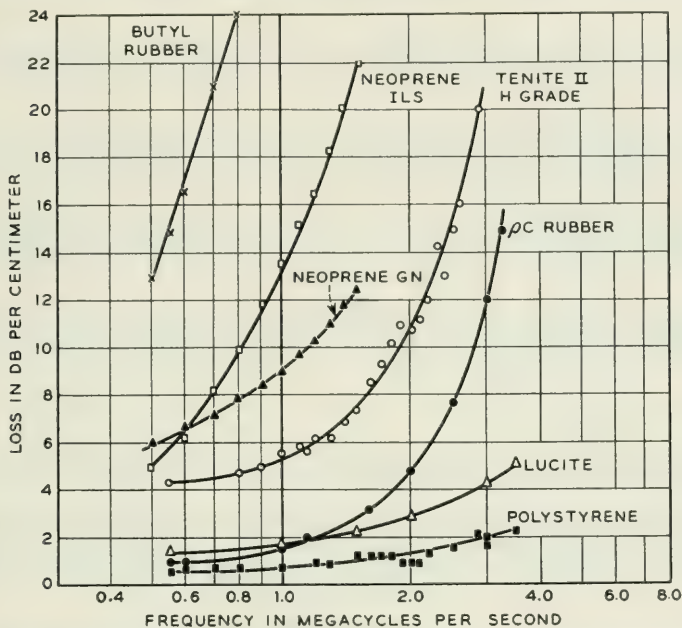


Fig. 24—Normal loss in db per centimeter measured as a function of frequency for several rubbers and plastics.

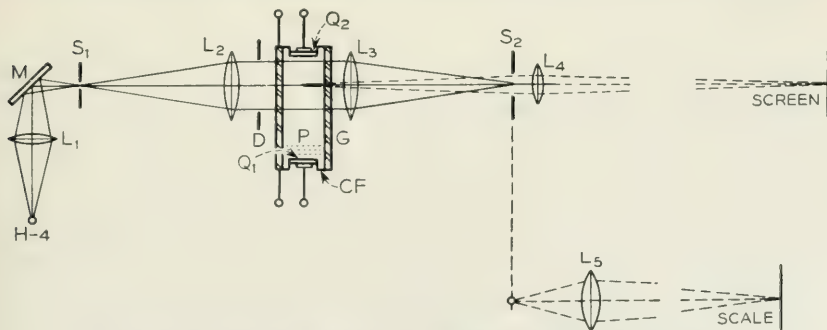


Fig. 25—Debye-Sears cell for making sound waves visible.

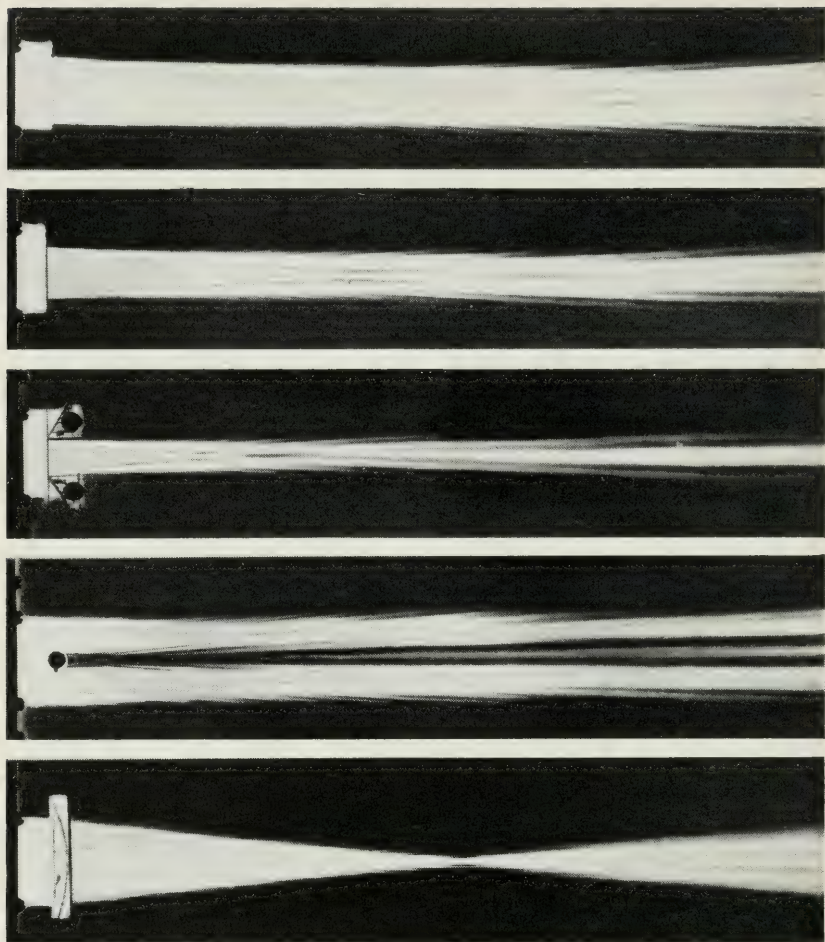


Fig. 26—Examples of refraction and focusing effects for sound waves.

and these have shown relaxations in such plastics as polyethylene and nylon 6-6. The simplest method for measuring one of the properties for longitudinal waves, i.e., the attenuation, is to measure the change in loss between two transducers in a liquid such as water, caused by inserting a sheet of the material. This process, used during the last war, results in the losses in db per centimeter for several rubbers and plastics, shown by Fig. 24. Indications of relaxation mechanisms are given by the rubbers and the plastic tenite II which is a cellulose acetate butyrate. The first fairly accurate method for measuring longitudinal sound

velocities¹² in plastics was the method of observing the focusing effect of a cylindrical lens made of the plastic. Sound waves can be made visible by the Debye-Sears technique of using a sound wave as a phase diffraction grating. Here light from a slit S_1 is made parallel by the lens L_2 and passes through the cell parallel to the wave fronts of the sound waves as shown by Fig. 25. The compressed parts of the medium retard the light waves more than the rarefied parts do and hence the medium acts as a phase diffraction grating. If a second slit S_2 is used which is small enough to pass only the zero order, a light valve action is obtained which modulates the light according to the sound wave intensity. If now the lens L_5 is used which focuses on the median plane of the tank, a picture of the sound beam is obtained as shown on Fig. 26. The bottom figure shows the focusing effect of a plastic and from the focal distance d and the radius of curvature r of the lense, one can calculate the velocity in a plastic compared to the velocity in the water by the formula

$$v_p = v_w / \left(1 - \frac{r}{d}\right) \quad (11)$$

This method gives velocities good to from 2 to 5 per cent depending on the attenuation in the lens.

G. W. Willard¹³ has devised recently a more accurate method for measuring sound velocities as shown schematically by Fig. 27. Here a plastic to be measured is placed half way across the sound beam in the liquid and light is sent along the wave front occurring in both the plastic and the liquid. If the waves are in phase the retardation in the two light gratings, corresponding to sound propagation in both media, add up and for a slit selecting the zero order the darkest pattern occurs on the photographic plate. If the two waves are just out of phase, the retardation is reversed in the two media and the lightest part occurs. With this relation it can be shown¹³ that the spacing d of light and dark lines

¹² W. P. Mason, *Piezoelectric Crystals and Their Application to Ultrasonics*, D. Van Nostrand, 1950, p. 404. It was used in this country by G. W. Willard as early as 1940. It was also used in Germany by J. Schaefer "Eine Neue Method Zur Messung der Ultraschallwellen in Festkorpern." Diss Strassburg, 1942. By making the front surface part of a cylinder, Schaefer also measured the shear velocity in a solid.

¹³ G. W. Willard, *J. Acous. Soc. Am.*, **23**, Jan. 1951, pp. 83-94. The origin of this multiple path interference method goes back to the work of R. Bär (Helvetia Physica Physica Acta Bd 13 page 61 (1940)) who attached a piezoelectric crystal to a bar with a 45° end section and set up transverse and longitudinal waves, in the bar. These waves produced longitudinal waves in a surrounding liquid and by observing the interference pattern between them, the longitudinal and shear constants could be determined for an isotropic medium. Willard's method as described above is much more direct and is capable of higher accuracies.

is related to the wavelength in the liquid λ_l and the wavelength in the solid λ_s , by

$$\frac{1}{d} = \frac{1}{\lambda_l} - \frac{1}{\lambda_s} \quad (12)$$

This corresponds to a velocity in the solid compared to a velocity in the liquid given by

$$v_s = \frac{v_l}{1 - v_l/fd} \quad (13)$$

where f is the frequency. Fig. 28 shows a photograph of a series of lines in a transparent plastic and a transparent plastic in the form of a wedge. It is seen that beyond the edge of the plastic there is a dark interference band for each one in the transparent plastic. This phenomenon is caused by the refraction of the sound wave that has traversed the plastic and the dark lines are lines of equal phase of the two waves in the liquid. The angle of the dark lines is half the refraction angle. Hence the velocity can also be determined by counting the number of dark bands in the liquid beyond the plastic. This makes it possible to measure the veloci-

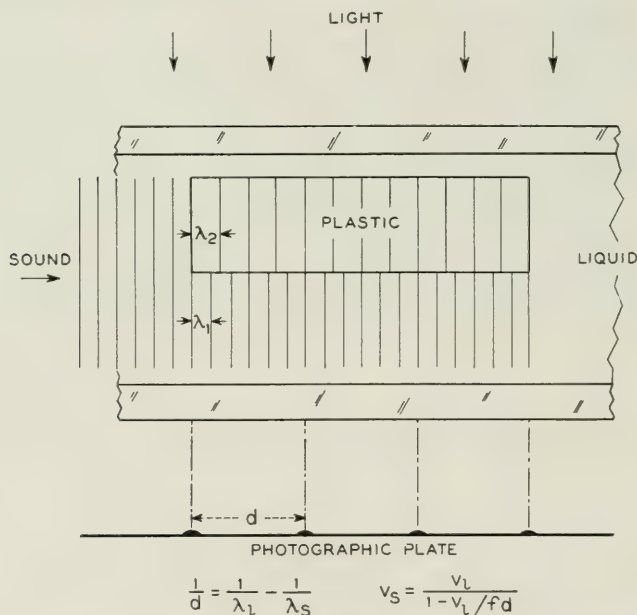


Fig. 27—Optical method for measuring sound velocities of plastics by comparing their velocity with that of a liquid such as water.

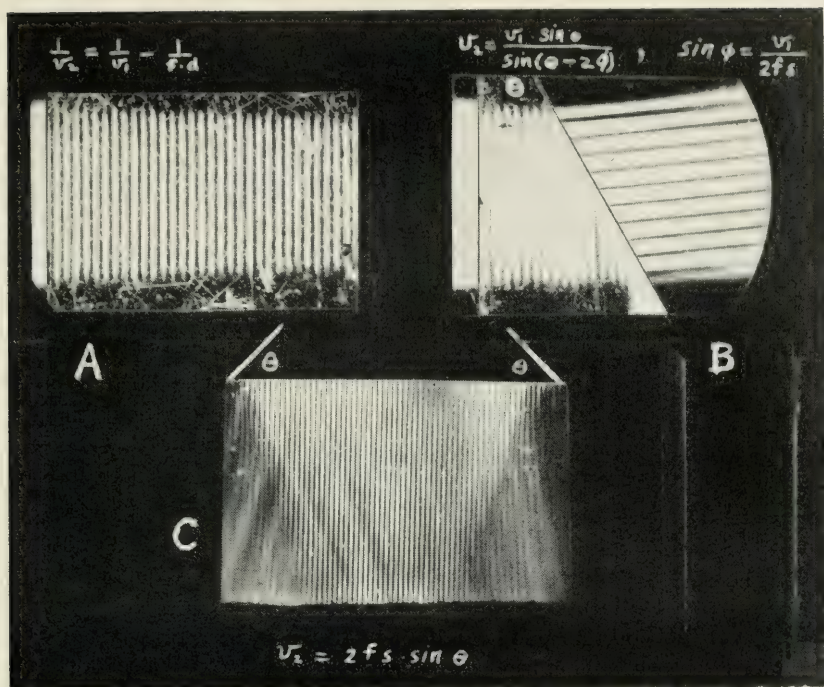


Fig. 28—Photographs of interference patterns from sound waves in liquids and plastics.

ties in opaque plastics. The accuracy of the method is better than 1 per cent if the attenuation is low enough to give a number of interference lines. For plastics of high internal loss, the method becomes somewhat inaccurate.

Typical measurements using this system are shown in Table V. Small changes in chemical composition and plasticizer content are shown up as can be seen from the table. Of particular interest is the difference between nylon 6-6 and polyethylene. Chemically as shown by Fig. 29, the two are identical except for the dipoles occurring for every 6 units of the ethylene chain. These dipoles have the effect of bonding adjacent layers together and result in a higher shearing modulus.

By attaching shear vibrating crystals to a right angled prism, as shown by the lower part of Fig. 28, with the direction of motion of the crystals parallel to the transmitting face, and setting up shear standing waves between the two crystals, the shear properties of the plastic can be measured. Longitudinal waves are generated in the liquid which interfere with one another and cause dark bands perpendicular to the plastic

TABLE V

Measured longitudinal velocity and attenuation at 25°C and 2.5 mc.
Longitudinal attenuation in DB/cm at 25°C and 2.5 mc except as noted.

Material	Long velocity $\times 10^{-5}$ cm/sec	Shear velocity 10^{-5} cm/sec	A DB/cm	Density
Dural, 17 ST.....	6.5	3.12	—	2.7
Brass, half hard.....	4.7	2.11	—	—
Polystyrene.....	2.35	1.12	2	1.05
Plexiglas.....	2.68		5	1.18
Tenite II, (cellulose acetate butyrate), 2% plasticizer.....	2.08		9	1.23
Tenite II, 13% plasticizer.....	2.02		10	1.21
Polyvinyl formal.....	2.68		10	1.24
Polyvinylidene chloride.....	2.4		18	1.71
Poly N-butyl methacrylate.....	1.96		5	1.05
Poly I-butyl methacrylate.....	2.08		6	1.05
Neoprene.....	1.51		20	.99
Polyethylene.....	2.0		4.7 F ^{1.11}	.90
Nylon 6-6 (3-30 megacycles).....	2.68		1.0 F ^{1.5}	1.11
Nylon 6-10 (3-30 megacycles).....	2.56		1.0 F ^{1.5}	1.11

surface. By determining the spacing of these lines the velocity of the shear waves can be determined.

Another method has also been developed which is more applicable for high loss materials. This is a pulsing method and is a modification of the method proposed by one of the authors for measuring the properties of small crystal specimens.¹⁴ Here longitudinal or shear crystals are soldered to the fused quartz rod as shown by Fig. 30 and a sample to be measured is placed between these by means of a liquid such as polyisobutylene which has a high shear elasticity. If the specimen has a small attenuation, this can be measured by taking the difference in the amplitude of successive reflections. If the specimen has a high loss, this does not work and another method has been used which consists in sending a pulse from both crystals.¹⁵ One crystal is then used to receive and it receives the wave sent through the sample and the wave reflected from the fused quartz-sample interface. By adjusting the amplitude until these two are equal and the frequency or phase of one channel until the waves cancel, a ratio of amplitudes and a frequency of half wavelength are accurately determined. From these the velocity and attenuation can be calculated.

This method has been applied to measuring the longitudinal and shear velocities of polyethylene and 6-6 nylon. The polyethylene was of "equilibrium" crystallinity and average molecular weight corresponding to

¹⁴ H. J. McSkimin, "Ultrasonic Measurement Techniques Applicable to Small Solid Specimens," *J. Acoust. Soc. Am.*, **22**, No. 4, July 1950, pp. 413-418.

¹⁵ H. J. McSkimin, *J. Acoust. Soc. Am.*, **23**, No. 4, pp. 429-435.

an intrinsic viscosity in xylene of $[\eta] = 0.89$ at 85°C . Fig. 31 shows the longitudinal velocity of polyethylene plotted as a function of frequency and temperature. The velocity rises with frequency and a dispersion is indicated. This is confirmed by the attenuation per wavelength curve for two different frequencies plotted as a function of temperature, Fig. 32. A definite dispersion is seen to occur with an activation energy of about 12 ± 2 kilocalories per mole. This could occur in either the λ constant or the shearing constants μ , but the data of Figs. 33 and 34 show definitely that it occurs in the shear component. Fig. 33 shows the shear velocity for four temperatures plotted as a function of frequency. This can be fitted for 30°C with a single relaxation mechanism having a relaxation frequency of 8 megacycles. To agree with the measured attenuation and velocity, there has to be a spreading of the single relaxation over a range as also occurs in liquids. The indicated shear stiffness below this relaxation frequency is 2.6×10^9 dynes/cm². Some

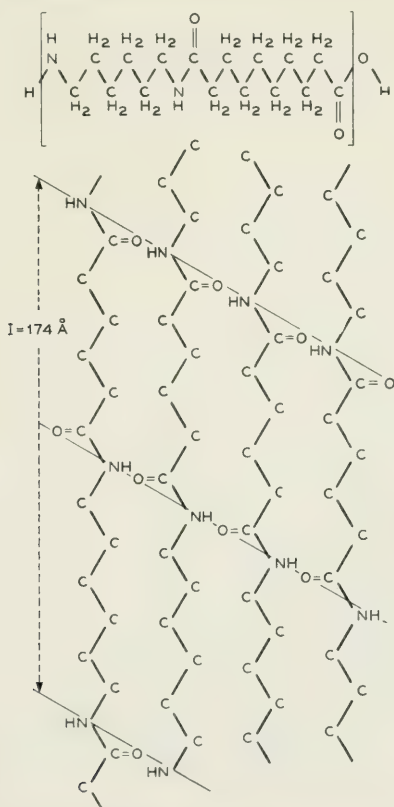


Fig. 29—Spatial structure of nylon 6-6.

data on the zero frequency shear modulus is obtained from the Young's modulus for a static pull which is from 30,000 to 50,000 pounds/square inch. Since the Young's modulus is three times the shearing modulus, the zero frequency shearing modulus should not exceed 1.1×10^9 dynes/cm². Hence one may expect that other relaxations will occur at lower frequencies.

Fig. 34 shows the attenuation per wavelength for shear waves. The solid line for 30°C represents the calculated attenuation per wavelength for the model assumed. If all the dissipation were due to shear mechanisms, the calculated attenuations would occur as shown by the 30°C

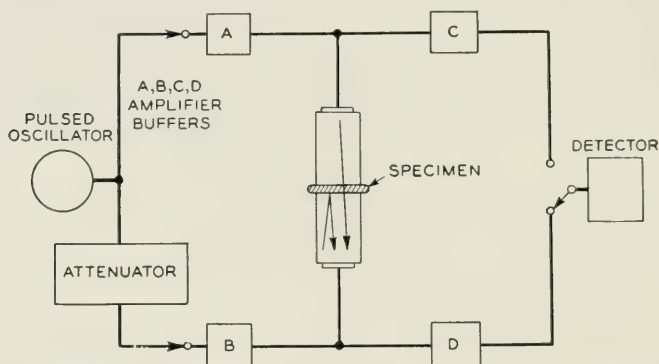


Fig. 30—Ultrasonic pulse method for measuring the velocities and attenuations of highly attenuating plastics.

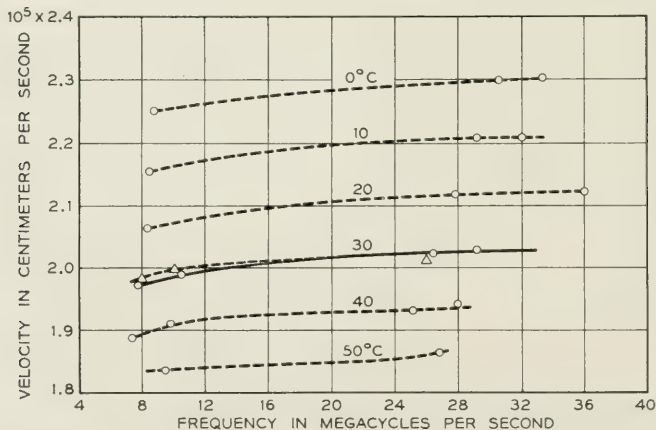


Fig. 31—Velocity of longitudinal waves in polyethylene plotted as a function of temperature and frequency.

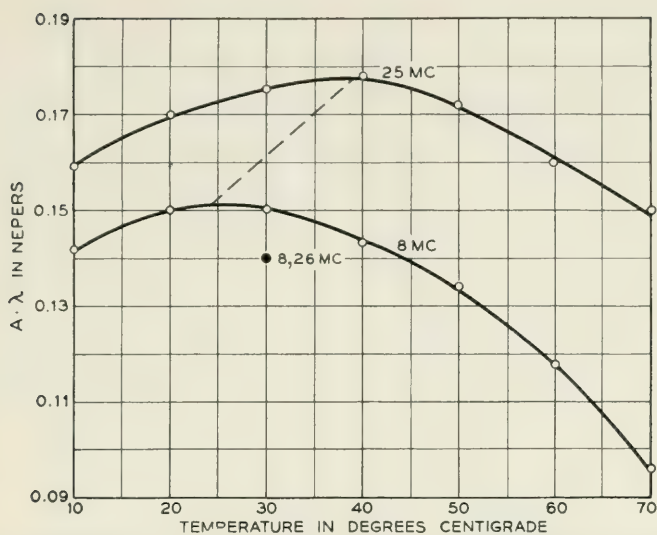


Fig. 32—Attenuation per wavelength for longitudinal waves in polyethylene plotted as a function of temperature and frequency.

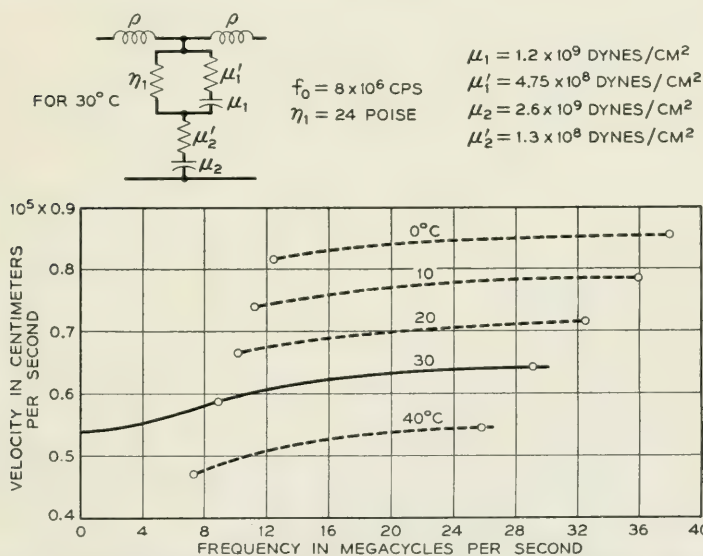


Fig. 33—Shear velocity of polyethylene as a function of frequency and temperature. Equivalent circuit shows elements necessary to account for the velocity and attenuation changes at 30°C.

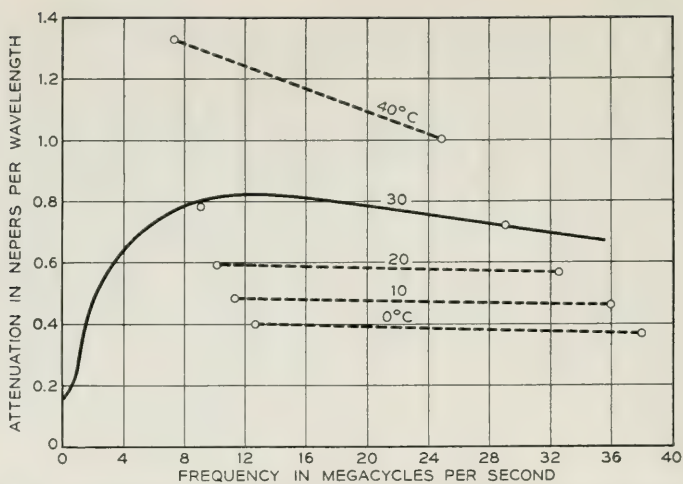


Fig. 34—Attenuation per wavelength for shear waves in polyethylene.

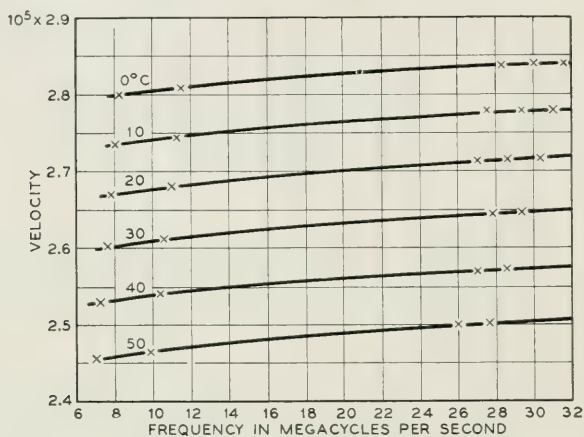


Fig. 35—Velocity of longitudinal waves in nylon 6-6 plotted as a function of temperature and frequency.

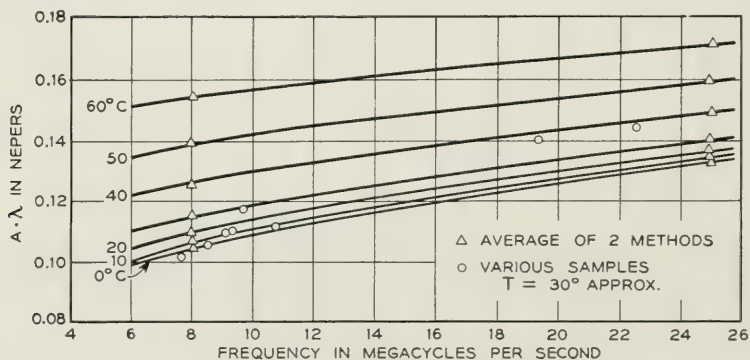


Fig. 36—Attenuation per wavelength for longitudinal waves in nylon 6-6 plotted as a function of temperature and frequency.

points of Fig. 32. Most of the loss is accounted for by shear mechanisms, but it appears that some compressional mechanisms may also be present.

The mechanism causing the relaxation in the megacycle range for polyethylene appears to be the same as for polyisobutylene, namely the relaxation of the shortest chain segment that is free to move. The chain segment acting appears to be longer than six chain units for similar measurements of nylon 6-6 show no relaxations in this frequency range. Fig. 35 shows the longitudinal velocity and Fig. 36 the attenuation per wavelength for longitudinal waves. Since the attenuation per wavelength is still increasing for nylon 6-6 at 25 megacycles a still shorter chain segment may be operating for this material. The shear velocity and attenuation per wavelength for nylon 6-6 are shown by Figs. 37 and 38.

Fig. 39 shows the shear stiffness of polyethylene and nylon 6-6 plotted

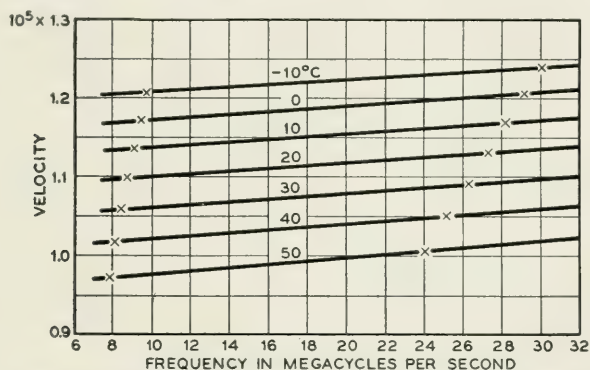


Fig. 37—Velocity of shear waves in nylon 6-6 plotted as a function of temperature and frequency.

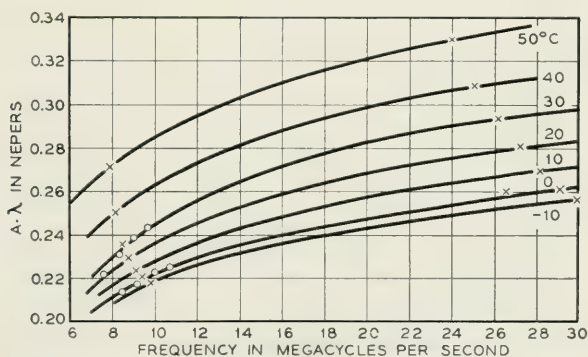


Fig. 38—Attenuation per wavelength for shear waves in nylon 6-6 plotted as a function of temperature and frequency.

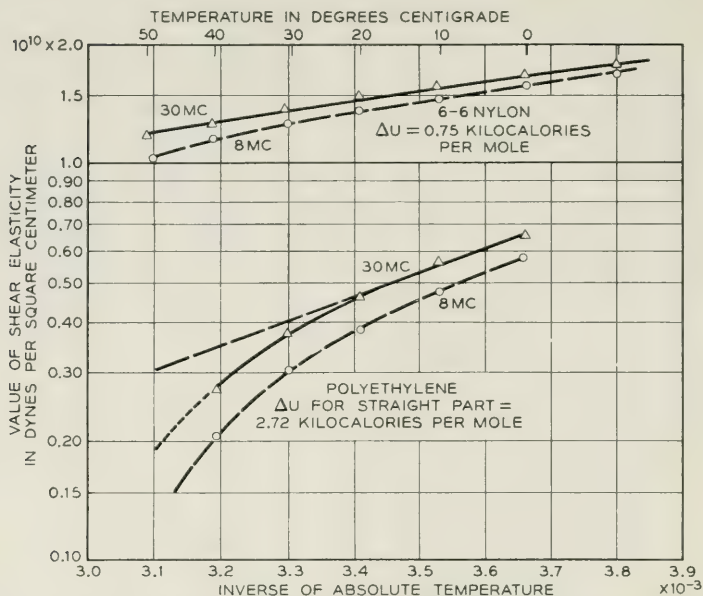


Fig. 39—Shear elasticity of polyethylene and nylon 6-6 plotted as a function of temperature and frequency.

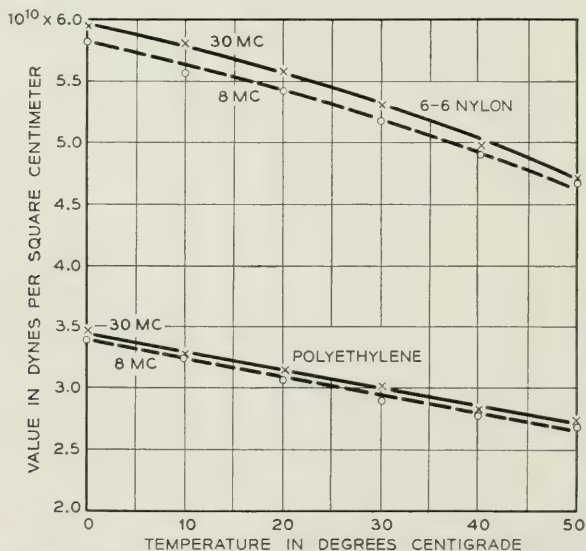


Fig. 40—Value of Lamé λ elastic constant for polyethylene and nylon 6-6 plotted as a function of frequency and temperature.

against $1/T$ where T is the absolute temperature. Both are plotted for 8 mc and 30 mc. The dispersion in both materials is evident. Below 30°C the shear elasticity of polyethylene varies exponentially with the temperature with an activation energy of 2.72 kilocalories per mole. Above this temperature a deviation occurs due to the approach to the melting temperature. Nylon has a smaller variation with temperature.

Comparing the longitudinal and shear wave measurements one can calculate the Lamé λ elastic constant and this is shown plotted on Fig. 40 for both polyethylene and nylon 6-6 as a function of temperature for

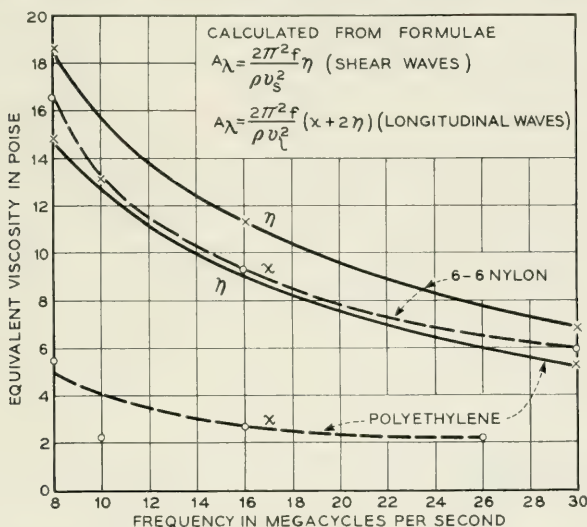


Fig. 41—Equivalent shear and compressional viscosities for polyethylene and nylon 6-6 plotted as a function of frequency for a temperature of 25°C .

two frequencies. The dispersion of λ for polyethylene is small but is more prominent in nylon 6-6. This correlates with the larger compressional viscosity component present for nylon 6-6 which as shown from Fig. 41 is as large as the shear viscosity. According to the structural rearrangement theory of compressional viscosity due to Debye,¹⁶ compressional viscosity can enter when some part of the chain can rearrange from one stable state to another stable state as a function of pressure. This rearrangement occurs across a potential barrier and hence requires a finite amount of time to occur. This lag in the rearrangement results in a compressional viscosity and as the frequency is increased, a frequency is found for which the motion can no longer occur in the time of a single

¹⁶ P. Debye, *Z. Elektrochem.*, **45**, 1939, p. 174.

cycle and the λ constant increases. It appears from these measurements that the dipole binding present in nylon 6-6 allows a greater structural rearrangement under pressure than can occur for polyethylene which has only linear chains.

VI. CONCLUSIONS

Measurements in dilute solutions, in pure polymer liquids and in non rigid solid polymers have all shown the presence of a shortest segment whose relaxation leads to a crystalline type of elasticity. In dilute polymer solutions the presence of a configurational type of relaxation and an entanglement relaxation of the shortest chain segment have been shown. For pure polymer liquids a quasi-configurational type of relaxation has been found for chain lengths greater than 60 segments, but for chain lengths less than 40 segments this type of relaxation disappears. From the difference between the high frequency shear elasticities measured for polyethylene and nylon 6-6 and the static measurement of Young's modulus, it appears that there may be other relaxations in these materials for lower frequency ranges.

For pure polyisobutylene and for nylon 6-6 there appear to be structural changes induced by pressure which account for a compressional viscosity and a dispersion in the λ elastic constant. This effect is smaller for polyethylene.

APPENDIX—EFFECT OF LIQUIDS ON THE PROPAGATION OF SHEAR WAVES IN RODS

For radially symmetric rods, the tangential particle displacement u_θ in the rod is given by

$$u_\theta = J_1(kr)e^{j\omega t - \theta z} \quad (1)$$

where

$$k^2 = \frac{\rho\omega^2}{\mu} + \theta^2 \quad (1A)$$

All other displacements are zero. In this equation waves are considered to be travelling in the $+z$ direction with a propagation constant $\theta = A + jB$, where A is the attenuation in nepers per cm and B the phase shift in radians per centimeter. μ is the shear stiffness which may be complex to take account of the dissipation within the rod.

From the defining relations for the stress strain equation

$$T_{r\theta} = \mu S_{r\theta} = \mu \left(\frac{\partial u_\theta}{\partial r} - \frac{u_\theta}{r} \right)$$

and the tangential particle velocity $\partial u_\theta / \partial t$, one can calculate the impedance Z per square cm. of cylindrical surface at $r = a$. This relation is

$$Z = \frac{-T_{r\theta}}{\dot{u}_\theta} = \frac{j\mu k}{\omega} \left[\frac{J_0(ka)}{J_1(ka)} - \frac{2}{ka} \right] \quad (2)$$

Since only the first mode is excited, parameters can be adjusted to keep k quite small, i.e. ($ka < .2$) and equation (2) can be simplified by using power series expansions for the Bessel functions. Neglecting higher order terms this results in

$$Z = \frac{-j\mu a k^2}{4\omega} \quad (3)$$

To evaluate the impedance of the liquid surrounding the rod, the torsional wave is first propagated along the length of the rod without the liquid, i.e. with $Z = 0$. Then from equation (3) $k = 0$ and from equation (1A)

$$\theta_0^2 = \frac{-\rho\omega^2}{\mu} = (A_0 + jB_0)^2 \quad (4)$$

where A_0 and B_0 are respectively the attenuation and phase shift in the rod alone. With the small loss in metal and glass rods A_0 can be taken equal to zero and

$$B_0 = \omega / \sqrt{\mu/\rho} = \omega/v_0 \quad (5)$$

where v_0 is the velocity of propagation in the rod alone.

When the liquid surrounds the rod, however,

$$k^2 = \frac{\rho\omega^2}{\mu} + \theta^2 = -(A_0 + jB_0)^2 + (A + jB)^2 \quad (6)$$

For the usual case where $(B + B_0) \gg (A + A_0)$, equation (6) approximates

$$k^2 = (B + B_0) (-\Delta B + j\Delta A) \doteq 2B_0 (-\Delta B + j\Delta A)$$

where ΔB is the increase in phase shift per centimeter and ΔA the increase in attenuation per cm, both directly measurable quantities. The

final working equation is then given by

$$Z = \frac{\mu a}{4\omega} \times 2B_0[\Delta A + j\Delta B] = \frac{\mu a}{2v_0} (\Delta A + j\Delta B) = \frac{\rho v_0 a}{2} (\Delta A + j\Delta B) \quad (7)$$

Since ΔA and ΔB are the attenuation and phase shift changes per unit length, then if l is the length of the rod, covered the total attenuation and phase shift changes will be ΔA and ΔB multiplied by $2l$. Hence if $\overline{\Delta A_0}$ and $\overline{\Delta B_0}$ are the measured attenuation and phase changes, the impedance Z becomes

$$Z = \frac{\rho v_0 a}{4l} (\overline{\Delta A_0} + j\overline{\Delta B_0}) \quad (8)$$

This derivation neglects the change of phase occurring at the intersection between the rod having no liquid and the rod surrounded by the liquid, but it can be shown that this is small and moreover, the change in the wave on leaving is equal and opposite to that occurring on entering and hence this correction cancels out. However, if the liquid is viscous enough there is a correction due to the fact that the measured impedance of equation (8) is for a cylindrical surface, whereas the desired impedance is the characteristic plane wave impedance. Obviously if the radius of curvature is sufficiently large no correction to equation (8) need be made. To obtain a suitable criterion one may consider waves propagated into the liquid from the surface of the rod and solve for the impedance per square cm. of the cylindrical surface. This neglects the variation with z , but since the wavelength along the rod is quite large, little error results from neglecting variations with z .

An outgoing cylindrical wave in the medium may be represented by

$$u_\theta = [J_1(kr)' - jY_1(kr)']e^{j\omega t} = H_1^{(2)}(kr)' e^{j\omega t} \quad (9)$$

where the primes refer to the wave in the liquid and

$$(k')^2 = \frac{\rho' \omega^2}{\mu'} = \frac{\rho'^2 \omega^2}{Z_k^2} \quad \text{or} \quad k'a = \frac{\rho' \omega a}{Z_k} \quad (10)$$

where $Z_k = \sqrt{\mu' \rho'}$ is the plane wave impedance of the liquid.

The shearing stress

$$T_{r\theta} = \mu' S_{r\theta} = \mu' \left[\frac{\partial u_\theta}{\partial r} - \frac{u_\theta}{r} \right],$$

and the tangential velocity may be obtained as before and the complex impedance over the cylindrical surface determined to be

$$Z = \frac{-T_{r\theta}}{\dot{u}_\theta} = -\mu' \left[\frac{\partial H_1^{(2)}(ka)'}{\partial r} - \frac{H_1^{(2)}(kr)'}{r} \right] / j\omega H_1^{(2)}(kr)' \quad (11)$$

Noting that for plane waves $Z_k = \sqrt{\rho'\mu'}$, one may eliminate μ' from (11) and evaluate Z at $r = a$ with the result

$$Z = j \left[\frac{H_0^{(2)}(ka)'}{H_1^{(2)}(ka)'} - \frac{2}{(ka)'} \right] Z_k \quad (12)$$

The above equation can be used to obtain a solution for Z_k in terms of the measured value of the cylindrical impedance Z for any set of parameters that may apply. Except for very heavy loading of the rod, however, the results of equation (8) may be used directly with little error. For example calculations indicate that for $|ka|' = 10$, the multiplier of Z_k of equation (12) is approximately $1.07 \angle -7.4^\circ$ for phase angle of $(ka)'$ of -25° . For $(ka)' < 10$, the correction multiplier rapidly becomes important. This same correction is applicable for the torsional crystal, but since this is only used for dynamic viscosities less than 10 poises, a correction is seldom necessary.

Relay Armature Rebound Analysis

BY ERIC EDEN SUMNER

(Manuscript received October 25, 1951)

Rebound of mechanical structures subsequent to impinging on stops generally has deleterious effects on their performance and should, therefore, be minimized. A considerable reduction in rebound can often be obtained by introducing additional degrees of freedom to the structure.

A mathematical treatise of the dynamics of rebound motion of systems representing idealized relay armatures is presented. Normalized differential equations of motion and their solutions for the "free" and "impact" intervals are derived for systems having one, two, and three degrees of freedom, allowing the rebound behavior of a specific system to be calculated. The equations of series of rebounds, and possible combinations of such series are considered next for systems having one and two degrees of freedom. The field of possible rebound maxima is mapped for a practical range of mass distribution constants, coefficients of restitution, and force ratios. A sufficiently broad optimum design region is indicated.

The results of this analysis have been checked closely on a model and have led to appreciable reduction of armature rebound in relay designs.

I. INTRODUCTION

In numerous types of mechanisms it is desirable to arrest the motion of a member at a particular point and to maintain it in this position. One of the simplest means of accomplishing this is to allow the moving member to impinge on a fixed member (stop) and to provide forces to tension it against this stop. Because the member to be arrested possesses kinetic energy and because the stop cannot generally absorb all of this energy, the moving member will rebound from the stop. The rebound motion generally deteriorates the performance of the mechanism and should be minimized.

Investigation of this phenomenon has been stimulated by the armature rebound problem in relay operation, where rebound from the front stop* tends to reclose contacts and must therefore be compensated for by additional (waste) travel, resulting in deleterious effects on speed and

* Among relay designers the front stop has been generally referred to as "back-stop". In this paper the terms front stop and heel stop have been used throughout for easier identification.

magnetic characteristics. Analysis in this paper will be directed towards relay armature systems, but it is also applicable to rebound in similar mechanisms.

II. STATEMENT OF PROBLEM

Analysis will be restricted to planar motion of armature systems having one, two, and three degrees of freedom as depicted in Figs. 1, 2, and 3. Generally one stop must be provided for each degree of freedom, although in the three-degree-of-freedom system of Fig. 3, two of the stops have been combined.

Applied forces F_1 , F_2 , F_3 , have been chosen so as to be most easily correlated with actual relay designs.

The initial condition in all cases will be a pure rotation about the "heel" just prior to a "zero" impact at the "front" of the armature. The "zero" impact will be followed by rebound motion and impacts at the various stops eventually bringing the armature to rest. The object

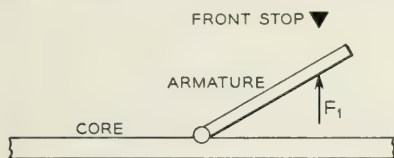


Fig. 1—Solidly hinged armature—one degree of freedom.

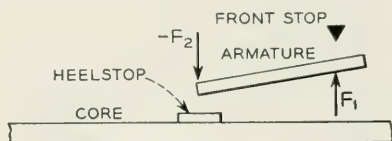
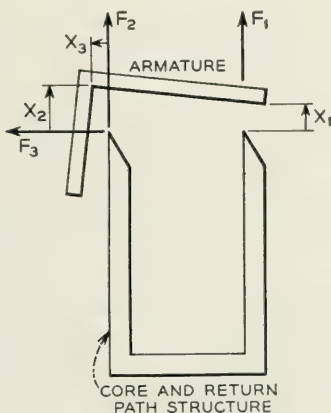
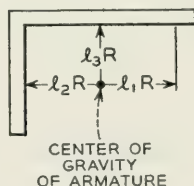


Fig. 2—Loosely hinged armature—two degrees of freedom.



(a)



(b)

Fig. 3—Armature system—three degrees of freedom.

will be to minimize rebound motion at the front, since this is usually near the point actuating the relay contacts.

The basic problem is then to find the response of the armature subject to aperiodic but well defined impulses, which are functions of the positions and velocities of the system.

III. ASSUMPTIONS

In order to facilitate the solution of this problem, the following modifying assumptions are made:

(1) As mentioned in the previous section, analysis is restricted to planar motion.

(2) The armature is assumed to be a rigid body.

(3) Stops are assumed to be very stiff, massless springs capable of energy absorption during impact with the armature. The associated coefficient of restitution is assumed constant. Core and stop vibration are neglected.

(4) The tensioning forces F_1 , F_2 , F_3 are assumed to be constant forces. (This is fairly closely true for moderate rebound amplitudes of practical relay structures.)

(5) All displacements are small relative to the dimensions of the system and in particular the angular displacement θ is sufficiently small so that

$$\cos \theta \doteq 1$$

$$\sin \theta \doteq \theta$$

IV. DERIVATION OF EQUATIONS OF MOTION

The derivation of the equations of motion resolves itself into the solution of two different types of intervals:

(1) *Free Interval*: This is the period during which the armature is not in contact with any of its stops and only the tensioning forces are acting.

(2) *Impact Interval*: During such intervals the armature is in contact with at least one of the stops. The stiffness of the latter is assumed so high that the tensioning forces during this interval may be neglected.

The three-degree-of-freedom case will be considered first and the others subsequently deduced from it by allowing some of the constants to approach zero.

A. Free Interval

The motion of the armature will be described by the displacement at the stop points: x_1, x_2, x_3 .* Let m be the mass and R the radius of gyration of the armature about the center of gravity. The latter is located by the dimensions l_1R, l_2R , and l_3R relative to the stop points, i.e., the points on the armature which contact the stops in the rest position (Fig. 3).

The equations of motion are derived in Appendix I and are put into dimensionless form:

$$\left. \begin{aligned} y_1 &= \frac{1}{2} \left[C_{11} + C_{12} \frac{(F_2)}{(F_1)} + C_{13} \frac{(F_3)}{(F_1)} \right] \left(\frac{t}{\tau} \right)^2 + \dot{y}_{10} \left(\frac{t}{\tau} \right) + \dot{y}_{10} \\ y_2 &= \frac{1}{2} \left[C_{21} + C_{22} \frac{(F_2)}{(F_1)} + C_{23} \frac{(F_3)}{(F_1)} \right] \left(\frac{t}{\tau} \right)^2 + \dot{y}_{20} \left(\frac{t}{\tau} \right) + \dot{y}_{20} \\ y_3 &= \frac{1}{2} \left[C_{31} + C_{32} \frac{(F_2)}{(F_1)} + C_{33} \frac{(F_3)}{(F_1)} \right] \left(\frac{t}{\tau} \right)^2 + \dot{y}_{30} \left(\frac{t}{\tau} \right) + \dot{y}_{30} \end{aligned} \right\} \quad (1)$$

where:

$$\left. \begin{aligned} y_i &= \frac{x_i}{\dot{x}_a \tau} = \frac{F_1}{\dot{x}_a^2 m} x_i \quad \dot{y}_i = \frac{d}{d\left(\frac{t}{\tau}\right)} y_i = \frac{\dot{x}_i}{\dot{x}_a} \\ \tau &= \frac{\dot{x}_a m}{F_1} \end{aligned} \right\} \quad (2)$$

\dot{x}_a is the front velocity \dot{x}_1 , just prior to the "zero" impact, and

$$\left. \begin{aligned} C_{11} &= (l_1^2 + 1) & C_{13} &= C_{31} = l_1 l_3 \\ C_{22} &= (l_2^2 + 1) & C_{12} &= C_{21} = (1 - l_1 l_2) \\ C_{23} &= (l_2^2 + 1) & C_{23} &= C_{32} = -l_2 l_3 \end{aligned} \right\} \quad (3)$$

$\dot{y}_{10}, \dot{y}_{20}, \dot{y}_{30}$, are the initial velocities and y_{10}, y_{20}, y_{30} the initial displacements for the free interval in question.

The equations of motion for a two-degree-of-freedom system are obtained, if $F_3 = 0$. Then for the two coordinates of interest:

$$\left. \begin{aligned} y_1 &= \frac{1}{2} \left[C_{11} + C_{12} \frac{(F_2)}{(F_1)} \right] \left(\frac{t}{\tau} \right)^2 + \dot{y}_{10} \left(\frac{t}{\tau} \right) + y_{10} \\ y_2 &= \frac{1}{2} \left[C_{21} + C_{22} \frac{(F_2)}{(F_1)} \right] \left(\frac{t}{\tau} \right)^2 + \dot{y}_{20} \left(\frac{t}{\tau} \right) + y_{20} \end{aligned} \right\} \quad (4)$$

* A summary of all notations used in this paper is given in Appendix IV.

For a one-degree-of-freedom system $\ddot{y}_2 = C_{21} + C_{22} \left(\frac{F_2}{F_1} \right) = 0$, whence

$$y_1 = \frac{1}{2} \left[C_{11} - \frac{C_{12}^2}{C_{22}} \right] \left(\frac{t}{\tau} \right)^2 + \dot{y}_{10} \left(\frac{t}{\tau} \right) + y_{10} \quad (5)$$

B. Impact Interval

The change of velocity at point "i" due to an impact at "i" is, by definition of the coefficient of restitution "k",

$$\Delta \dot{x}_1 = - (1 + k_i) \dot{x}_i$$

It is assumed here that the action of the stops are true impacts, i.e., the changes in velocity take place while there is negligible motion of the body. The velocity changes then occur as instantaneous rotation about the conjugate axis, leading to the general relation for an impact at point "i":

$$\dot{y}_{j0n} = \dot{y}_{je(n-1)} + K_{ji} \dot{y}_{ie(n-1)} \quad (6)$$

The first subscript indicates the coordinate, the second subscript indicates the beginning (0) or the end (e) of the free interval described by the third subscript. The impact transfer coefficient K_{ji} relating a velocity change at point "j" to an impact at point "i":

$$K_{ji} = - \frac{C_{ji}}{C_{ii}} (1 + k_i) \quad (7)$$

Equations (1) through (7) allow any one specific case to be mapped, if the mass distribution and force ratio are known. A sample of such mapping of rebound motion for a rectangular two-degree-of-freedom armature appears in Fig. 4.

V. ANALYSIS OF REBOUND PATTERN—ONE-DEGREE-OF-FREEDOM SYSTEM

The rebound pattern for the one-degree-of-freedom system—as derived in Appendix II—consists of an infinite series of parabolic arcs of diminishing amplitudes. The structure comes to rest after a finite time interval. The maximum rebound occurs during the first bounce and equals

$$Y = - \frac{k^2}{2C} \quad (8)$$

where

$$C = C_{11} - \frac{C_{12}^2}{C_{22}} \quad (9)$$

The system returns to rest at

$$\frac{t}{\tau} = \frac{2}{C(1-k)} \quad (10)$$

VI. ANALYSIS OF REBOUND PATTERN—TWO-DEGREE-OF-FREEDOM SYSTEM

The reason for choosing a two-degree-of-freedom system over a one-degree-of-freedom system would be, in keeping with the philosophy of

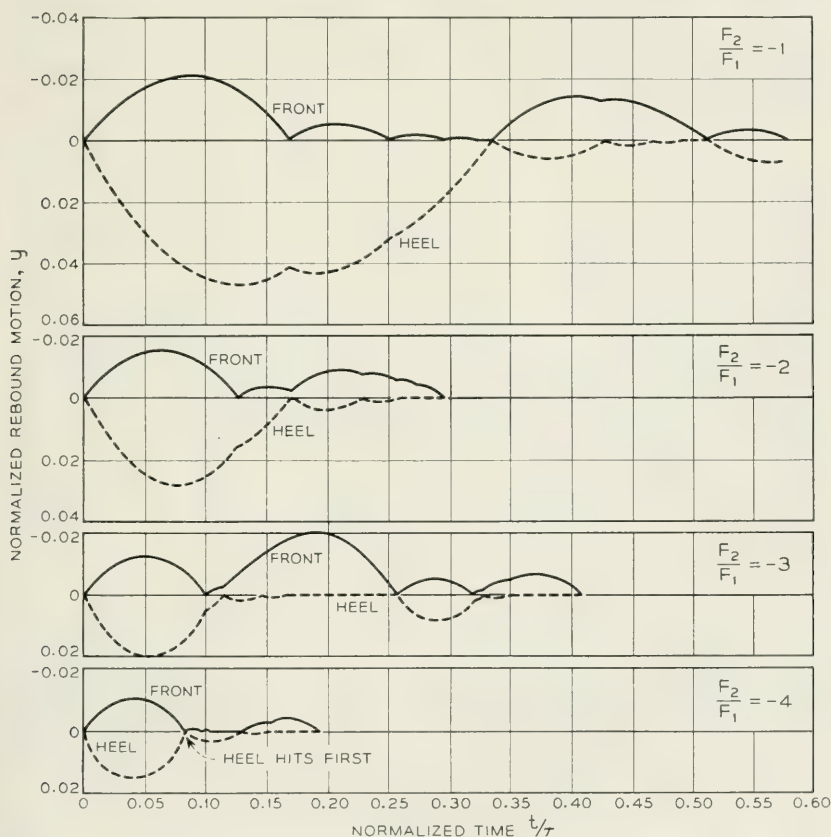


Fig. 4—Front and heel motion of plate type armature.

this treatment, to reduce Y_1 , the greatest excursion at the front. In order to simplify mapping, this maximum excursion will be expressed as $2CY_1$, the ratio of Y_1 to Y as given by Equation (8) for the case of $k = 1$. Thus $2CY_1$ is the ratio of the greatest excursion of the two-degree-of-freedom system under consideration to the greatest excursion of the corresponding perfectly elastic one-degree-of freedom system.

We first introduce two basic constants which are functions of the mass distribution relative to the stop locations:

$$M_{ij} = \frac{C_{ij}^2}{C_{ii}C_{jj}} \quad (11)$$

This constant represents a mechanical coupling coefficient. As $M_{ij} = M_{ji}$, the two-degree-of-freedom system under consideration here has only one such non-trivial constant M_{12} .

The second constant represents a force transformation factor from the "j" coordinate to the "i" coordinate:

$$P_{ij} = \frac{C_{ij}}{C_{ii}} \quad (12)$$

In the analysis of the two-degree-of-freedom system only P_{12} is important.

If there is to be any heel motion, the "zero" impact at the front must impart a positive velocity to the heel. By Equations (6), (7), and (12), this requires that P_{12} be negative, which in turn implies that $k_1 k_2 > 1$. For the limiting case of $k_1 k_2 = 1$, $P_{12} = M_{12} = 0$ and no coupling exists between the heel and the front. Physically this means that the two stops are the centers of percussion of each other and the system will act as a simple hinge.

With the above foundation, it is possible to analyze the patterns of motion and maximum rebound amplitudes.

A. Motion Immediately Following "Zero" Impact

After the "zero" impact at the front, both front and heel will lift off in accordance with impact Equation (6) and continue to move in accordance with the free interval Equations (4). Whether the next impact occurs at the front or the heel depends on their respective periods, t_1 and t_2 :

$$\frac{t_1}{t_2} = \frac{1 + \frac{P_{12}}{M_{12}} f}{1 + P_{12} f} \frac{k_1}{1 + k_1} \quad (13)$$

where:

$$f = \frac{F_2}{F_1}$$

A large value of t_1/t_2 will result in a series of heel impacts and the heel will come to rest while the front is still displaced from the stop. This will be called a complete heel series. A small value of t_1/t_2 results in a similar complete front series. If t_1/t_2 is near unity, a limited number of impacts on one end are followed by an impact on the other end, etc. An analysis of front and heel series follows:

B. Front Series

If $t_1/t_2 < 1$ a series of front impacts occurs. The impact velocities at the front are

$$\dot{y}_{1n} = 1, k, k^2, \dots, k^n \quad (14)$$

The corresponding time intervals are

$$T_{1n} = \frac{2k_1}{A}, \frac{2k_1^2}{A}, \dots, \frac{2k_1^n}{A} \quad (15)$$

where

$$A = (C_{11} + C_{12}f)$$

During this time, the heel velocity and displacement are given by

$$\left. \begin{aligned} \dot{y}_{20(n+1)} &= \dot{y}_{20n} + \left[\frac{2B}{A} - P_{12}(1 + k_1) \right] y_{x0n} \\ y_{20(n+1)} &= y_{20n} + \left[\frac{2B}{A} \dot{y}_{1n} + \dot{y}_{20n} \right] \dot{y}_{10n} \end{aligned} \right\} \quad (16)$$

where

$$B = C_{12} + C_{22}f$$

The velocity and displacement at the heel after a given number of front impacts are obtained by a summation of Equations (16). For a complete front series $n \rightarrow \infty$, and

$$\left. \begin{aligned} y_{2\infty} &= \frac{2k_1}{A(1 + k_1)^2} \left[\frac{Bk_1}{A} - P_{12} \right] \\ \dot{y}_{2\infty} &= \frac{1}{1 - k_1} \left[\frac{2Bk_1}{A} - P_{12}(1 + k_1) \right] \end{aligned} \right\} \quad (17)$$

In addition it is useful to set down energy equations in order to simplify evaluation of greatest rebound for the various groups of rebound patterns. The kinetic energy function T is evaluated in Appendix I. A potential energy term V —the work done against F_1 and F_2 from the equilibrium position—is introduced. If T_0 is the total energy of the system prior to the “zero” impact, then

$$\frac{T + V}{T_0} = \dot{y}_1^2 + \frac{M_{12}}{P_{12}^2} \dot{y}_2^2 - \frac{2M_{12}}{P_{12}} \dot{y}_1 \dot{y}_2 - 2C(y_1 + fy_2) \quad (18)$$

The energy loss due to n front impacts is

$$-\Delta \left(\frac{T + V}{T_0} \right) = (1 - k_1^{2n}) (1 - M_{12}) \dot{y}_{1e0}^2 \quad (19)$$

For a complete front series $n \rightarrow \infty$, and

$$-\Delta \left(\frac{T + V}{T_0} \right) = (1 - M_{12}) \dot{y}_{1e0}^2 \quad (20)$$

If a complete front series follows the “zero” impact, $\dot{y}_{1e0} = 1$ and

$$-\Delta \left(\frac{T + V}{T_0} \right) = (1 - M_{12}) \quad (21)$$

After completion of this “initial” front series, the system maintains only one degree of freedom (rotation about the front) until a heel impact occurs. By setting $\dot{y}_1 = y_1 = y_2 = 0$ in (21) we obtain the heel impact approach velocity $\dot{y}_2 = P_{12}$.

Apparently energy loss due to n front impacts is a function of M_{12} , k_1 , and the approach velocity of the first impact.

C. Heel Series

An analysis similar to the above can be made for partial and complete heel series following the “zero” impact. This is demonstrated in Appendix III, yielding, for $k_1 = k_2^*$

$$\left. \begin{aligned} y_{1e\infty} &= \frac{AP_{12}(1+k)^2}{B(1-k)^2} \left[\frac{AP_{12}}{B} - \frac{k(1-k)}{1+k} - M_{12}k \right] \\ \dot{y}_{1e\infty} &= \frac{1+k}{1-k} \left[\frac{2AP_{12}}{B} - \frac{k(1-k)}{(1+k)} - M_{12}(1+k) \right] \end{aligned} \right\} \quad (22)$$

* The more general form $k_1 \neq k_2$ can be obtained as indicated in Appendix III.

The energy relationships for heel series are

$$-\Delta \left(\frac{T + V}{T_0} \right) = (1 - k_2^{2n}) \frac{M_{12}(1 - M_{12})}{P_{12}^2} \dot{y}_{2e0}^2 \quad (23)$$

For a complete series $n \rightarrow \infty$, and

$$-\Delta \left(\frac{T + V}{T_0} \right) = \frac{M_{12}(1 - M_{12})}{P_{12}^2} \dot{y}_{2e0}^2 \quad (24)$$

If a complete heel series follows the "zero" impact, $\dot{y}_{2e0} = P_{12}(1 + k_1)$, and

$$-\Delta \left(\frac{T + V}{T_0} \right) = M_{12}(1 - M_{12})(1 + k_1)^2 \quad (25)$$

Finally, for the special case where a complete heel series follows an initial complete front series $\dot{y}_{2e0} = P_{12}$, and

$$-\Delta \left(\frac{T + V}{T_0} \right) = M_{12}(M_{12} - 1) \quad (26)$$

It is to be noted that the energy loss due to a partial heel series is a function of M_{12} , P_{12} , k_2 , and the approach velocity of the first impact, but that the equation for a complete heel series does not contain k_2 . Finally, a complete initial heel series is a function of only M_{12} and k_1 .

D. Complete Mapping of Problem

Equations (1) through (26) make it possible to completely map the two-degree-of-freedom rebound problem. The relative maximum amplitude $2CY_1$ and the rebound pattern will be determined.

Examination of the necessary equations, show that $2CY_1$ is in all cases a function of four parameters: k_1 , k_2 , M_{12} and $P_{12}f$. Of these, k_2 enters only if a partial heel series occurs prior to the time of maximum rebound. If it is assumed that for this limited group of cases $k_2 = k_1 = k$, the number of parameters is reduced to three: k , M_{12} , $P_{12}f$.

In Figs. 5 to 10, $2CY_1$ is plotted against $P_{12}f$ for the most useful range of $1/8 < M_{12} < 1/2.5$, $0.3 < k < 0.6$ and $0 < P_{12}f < 10$.

As $P_{12}f$ is increased from zero to infinity (corresponding to an increase in the heel tension F_2), the rebound pattern goes through some or all of five regions. The criterion for location in any one region is based upon the parameter

$$Q = \frac{1 + \frac{P_{12}}{M_{12}}f}{1 + P_{12}f} = \frac{t_1(1 + k)}{t_2 k} \quad (27)$$

Region I—Complete initial front series for $1 < Q < 1/k$. Within this region, if $M_{12}^2 > \frac{(1 - M_{12})k^2}{1 + P_{12}f}$, the maximum rebound occurs during the first bounce and

$$2CY_1 = \frac{(1 - M_{12})k^2}{1 + P_{12}f} \quad (28)$$

If the maximum rebound occurs later, it must occur during a complete heel series which follows the initial complete front series. From Equations (21) and (26)

$$2CY_1 = M_{12}^2 \quad (29)$$

By comparing Equations (28) and (29), the critical requirement

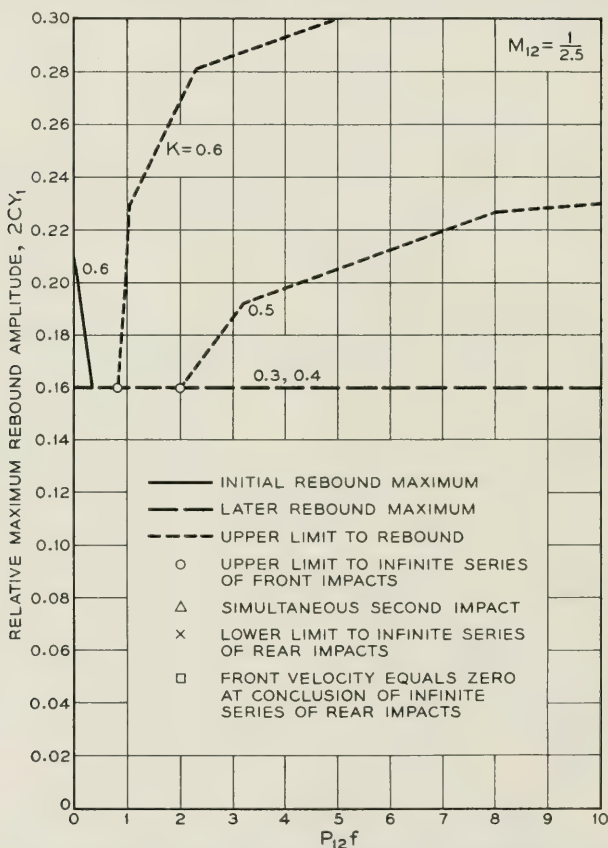


Fig. 5—Relative maximum rebound amplitude for $M_{12} = 1/2.5$.

for the latter case is that $P_{12}f > \frac{(1 - M_{12})}{M_{12}^2} k^2 - 1$. It should be noted that while the first rebound maximum, shown in solid lines on Figs. 5 to 10, is always realized, the later rebound given by (29) is an upper limit—shown in dashed lines—and is not always realized. In the dashed regions, phasing is extremely critical and small variations in the parameters may cause large variations in maximum rebound. From an engineering standpoint these regions are essentially undesirable.

Region II—Partial initial front series for

$$\frac{1}{k} < Q < \frac{1+k}{k}.$$

This region is one of critical phasing, and attention is limited to special cases leading to maximum rebound. These cases occur when a

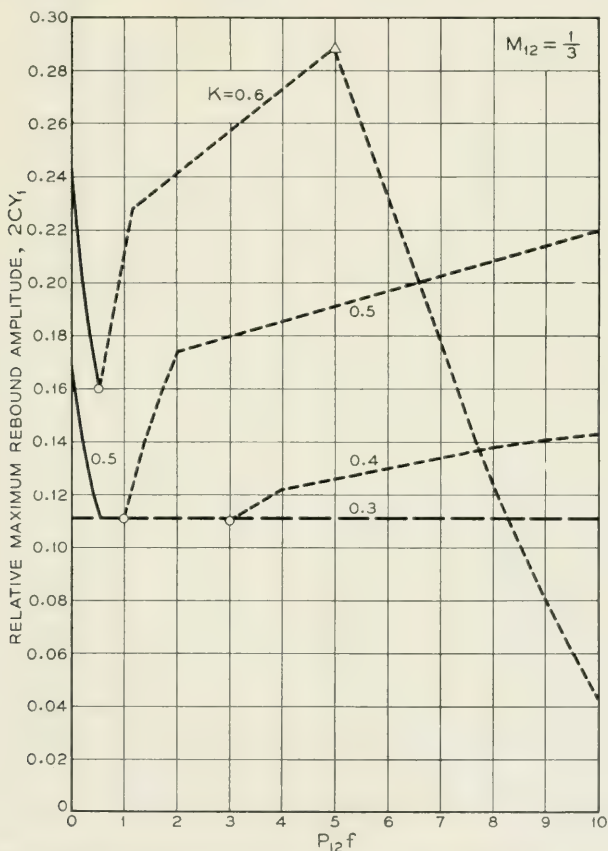


Fig. 6—Relative maximum rebound amplitude for $M_{12} = 1/3$.

heel impact immediately follows the last front impact of the series. These cases occur at

$$Q = \frac{1 - k^n}{k - k^n} \quad (30)$$

and lead to rebound amplitudes

$$2CY_1 = M_{12} + (1 - M_{12})[k^{2n} - M_{12}(1 - k^n)^2] \quad (31)$$

In Figs. 5 to 10, these special points are plotted and connected with straight dotted lines, which therefore indicate upper limits to rebound.

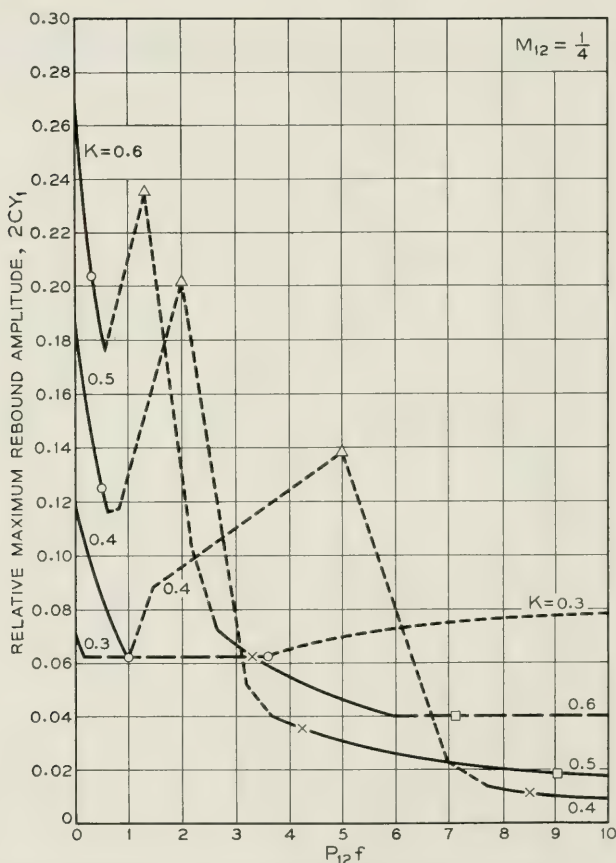


Fig. 7—Relative maximum rebound amplitude for $M_{12} = 1/4$.

Region III—Partial initial heel series for

$$\frac{1+k}{k} < Q < \frac{1+k}{k(1-k) + M_{12}k(1+k)}.$$

This is a region of critical phasing, and values were determined only for the maximum cases, where a front impact just precedes the last impact of the partial initial heel series. Here:

$$Q = \frac{1 - k^{n+1}}{\frac{k(1-k)}{1+k} + k(1-k^n)M_{12}} \quad (32)$$

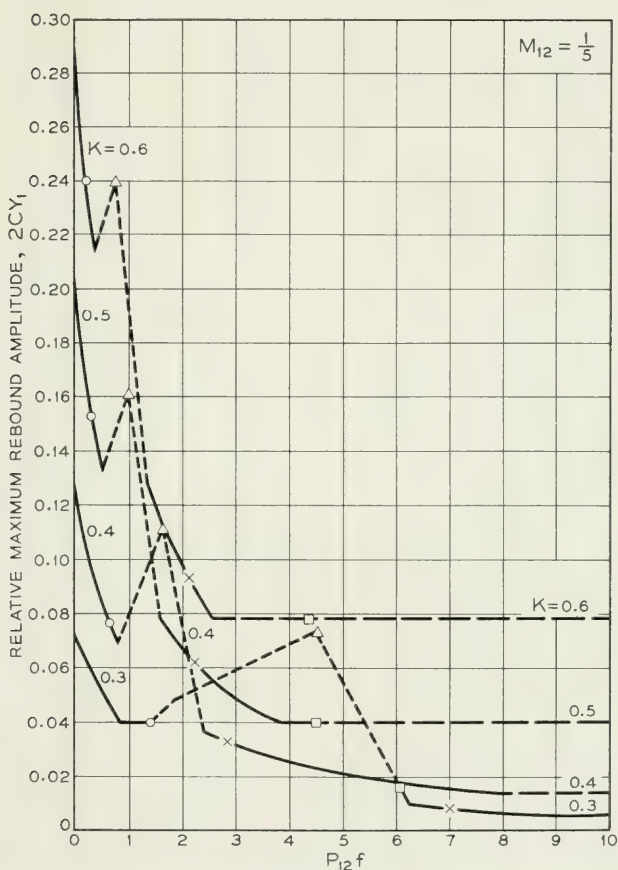


Fig. 8—Relative maximum rebound amplitude for $M_{12} = 1/5$.

and:

$$\begin{aligned}
 2CY_1 = 1 - (1 - M_{12})(1 - k^2) \{1 + [k - M_{12}(1 + k)(1 - k^n)]^2\} \\
 - M_{12}(1 - M_{12})(1 + k)^2 \{1 - k^{2n} \\
 + [k - k^n - M_{12}(1 + k)(1 - k^n)]^2\}
 \end{aligned} \quad (33)$$

Region IV—Complete initial heel series. A complete initial heel series implies that when the heel has come to rest, the front is still away from the stop. When the front finally meets its stop, the situation is identical with that just prior to the zero impact except that the energy content of the system is lower. The pattern must then repeat with diminished amplitudes. For this region we recognize two groups of cases. The first

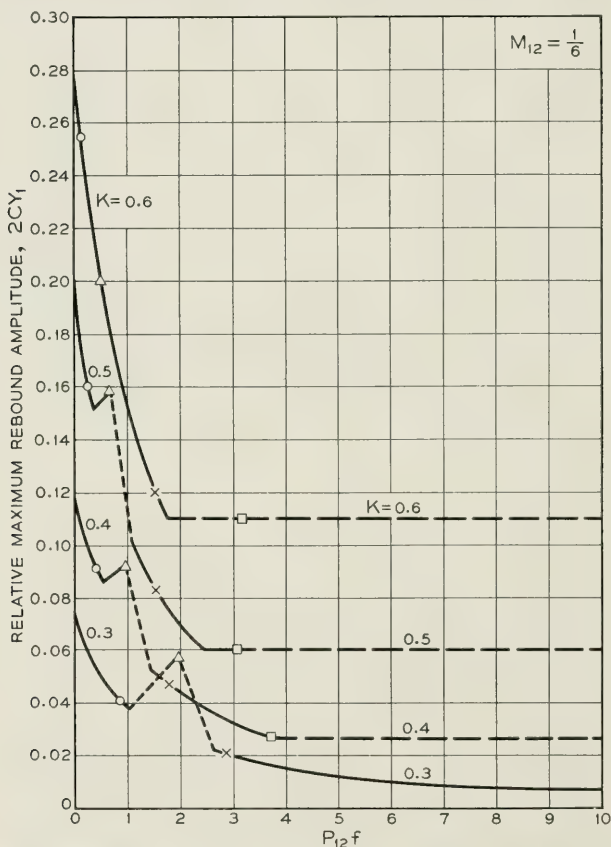


Fig. 9—Relative maximum rebound amplitude for $M_{12} = 1/6$.

group is that in which the front velocity is positive at the completion of the heel series. In that case

$$\frac{2(1+k)}{k(1-k) + M_{12}(1+k)^2} > Q > \frac{(1+k)}{k(1-k) + M_{12}k(1+k)^2}$$

and

$$2CY_1 = \frac{(1 - M_{12})k^2}{1 + P_{12}f} \quad (34)$$

For the second group the front velocity is still negative when the heel comes to rest from which point on the system acts as a one-degree-of-

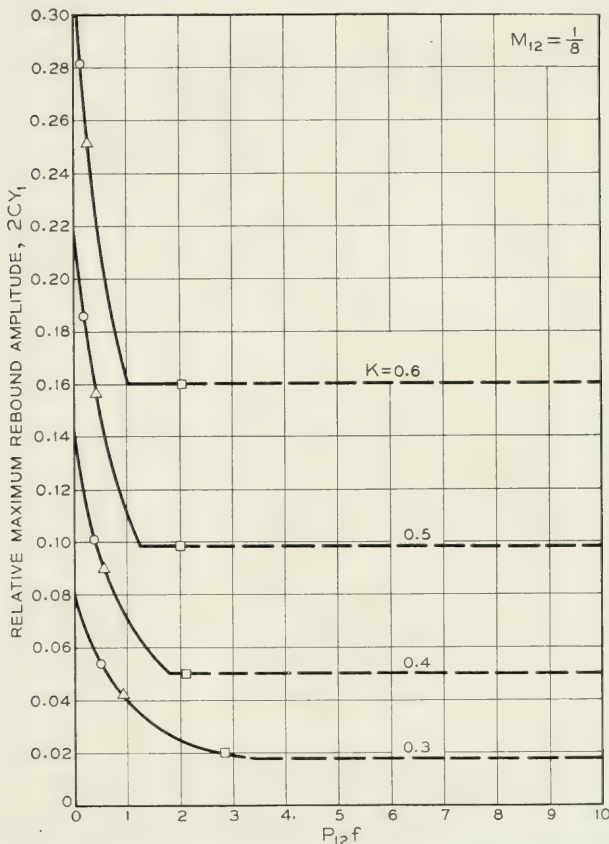


Fig. 10—Relative maximum rebound amplitude for $M_{12} = 1/8$.

freedom system until the next front impact. The requirement for this group is that

$$Q > \frac{2(1+k)}{k(1-k) + M_{12}(1+k)^2}$$

and the maximum rebound is given by

$$2CY_1 = M_{12} - (1 - M_{12})[(1 - k^2) + M_{12}(1 + k)^2] \quad (35)$$

It is to be noted that in the upper part of Group 1 the amplitude increases with successive heel impacts. This can be explored through the use of Equation (22). For simplicity of mapping, however, the limit given by Equation (35) has been extended back from the lower boundary of Group 2 until it intersects the line marking the first rebound amplitude of Group 1.

In Figs. 5 to 10 the respective regions have been identified by means of the symbols indicated below:

Region I	from $P_{12}f = 0$	to \circ
Region II	from \circ	to \triangle
Region III	from \triangle	to \times
Region IV, Group 1	from \times	to \square
Region IV, Group 2	from \square	to $P_{12}f \rightarrow \infty$

E. Discussion of Rebound Charts

Aside from quantitative data contained in Figs. 5 to 10, the following general trends are of interest:

For values of $M_{12} > \frac{1}{4}$, and the values of k under consideration, most of the useful range of $P_{12}f$ involves critical phasing and the rebound maxima are relatively high.

For values of $\frac{1}{6} \leq M_{12} \leq \frac{1}{4}$, consistently controllable rebound amplitude may be obtained.

For values of $M_{12} < \frac{1}{6}$ rebound increases again and the structure approaches the one-degree-of-freedom case.

VII. ANALYSIS OF REBOUND PATTERNS—THREE-DEGREES-OF-FREEDOM SYSTEM

Rebound pattern analysis as in Parts V and VI has so far not been performed for the three-degree-of-freedom system, partly because of complexity, and partly because for the system of Fig. 3 friction at the hinging stop will greatly influence the motion.

However, it is felt that the approach and notation of the analysis presented here is sufficiently general to allow extension of the rebound pattern analysis to the three-degree-of-freedom case. At any rate, with the assumption of the magnitudes of frictional forces, the basic equations of Part IV may be used to plot any particular case.

VIII. ARMATURE REBOUND MODEL

In order to verify the formal analysis presented in Parts III, IV and V, a large model of a two-degree-of-freedom system was constructed. It consisted essentially of a large bar constrained to move in a plane, biased against two stops, and to the ends of which writing pens were attached. As rebound conditions were simulated by releasing the bar against its stops, chart paper moved at right angles to the bar motion and thus produced a record of end displacement versus time.

By changing spring members and attaching masses to the bar, it was possible to vary the mass distribution and the biasing forces.

The results obtained closely agreed with those suggested by the analysis. The maximum rebound amplitudes were generally somewhat lower probably due to frictional effects.

IX. RELAY DESIGN CRITERIA RESULTING FROM ARMATURE REBOUND ANALYSIS

A. *Limitation of Analysis*

The assumptions which this analysis is subject to have been described under Part II. As applied to relays and probably the majority of mechanical structures, one assumption is most frequently and severely violated. The stops, which have been assumed to be stiff springs associated with a definite coefficient of restitution are, in practice, massive bodies which dissipate energy through excitation of high frequency modes of vibration. Accordingly, the assumption that the stops are at rest is violated, particularly if the mechanism is subject to repetitive (pulsing) impacts and the stop vibration does not decay greatly in the repetition period.

However, mechanisms designed in accordance with this analysis have performed well even under moderate pulsing conditions if the sensitive phasing region was avoided. In addition, every effort should be made to reduce the amount and duration of stop and mounting structure vibration by making them stiff, massive, and dissipative, if possible.

B. *Design Criteria*

1. Type of Armature Structure.

The selection of the number of degrees-of-freedom for an armature structure depends on the expected coefficient of rebound as well as practical considerations.

It can be shown without great difficulty that for very low coefficients of rebound the one-degree-of-freedom system is preferable. This is quite obvious when one considers the limiting value of $k = 0$. In this case the one-degree-of-freedom system will have no rebound whatsoever, while the two-degree-of-freedom system has a heel bounce followed by rebound at the front. The value of k below which the one degree system is preferable varies with the mass distribution relative to the stop points, being 0.18 for a rectangular plate armature with stops located at its edges.

Experience indicates that k in most practical relays and similar mechanical structures varies from 0.3 to 0.6. Hence the two-degree-of-freedom system is superior in all practical cases to the solidly hinged armature.

As far as three and higher degree-of-freedom systems are concerned, it may be said that generally the greater the number of modes resulting in impacts, the quicker the rebound energy can be diverted and dissipated and the lower theoretical rebound values can be obtained. This consideration would favor systems containing many degrees of freedom. However, while multi-degree-of-freedom systems can reach very low rebound values, their motion (phasing) must be very closely controlled or they may prove to be inferior to simpler systems particularly under vibratory (pulsing) operation. It is this difficulty which makes it appear that the two-degree-of-freedom system offers the best promise with the three-degree system also quite promising. By the same reasoning, additional spurious rocking modes should be minimized.

2. Armature Mass.

The armature mass should be as low as possible. This will minimize stop and structure vibration. In addition, in relay applications light armatures tend to increase magnetic "drag" losses of the armature during the release motion.

3. Stops and Mounting Structure.

As discussed before, it is desirable to reduce the amount and duration of stop and mounting structure vibration.

4. Coefficient of Restitution.

The coefficient of restitution should be kept low. Stops having low stiffness should, therefore, be avoided.

5. Biasing Forces.

F_1 should be kept as high as practicable.

For proper energy loss during impacts, the motion between impacts must occur outside the region of the compression, i.e., the armature and stop must separate. Therefore, because all practical stops have a finite stiffness, the biasing forces (F_1 , F_2 , etc.) should produce a static deflection less than say, arbitrarily, 5 per cent of the maximum expected rebound amplitude.

6. Design Parameters for Two-Degree-of-Freedom Systems.

As clearly indicated in Figs. 5 to 10 for the practical range of coefficients of restitution, most consistently good results are obtained with a coupling factor $M_{12} = \frac{1}{16}$ to $\frac{1}{4}$. This factor is most easily adjusted by correct placement of the front stop.

For the above range of M_{12} the force ratio F_2/F_1 should be such as to make the product

$$\begin{aligned} P_{12} \frac{F_2}{F_1} &> 4 & M_{12} &= \frac{1}{4} \\ &> 3 & M_{12} &= \frac{1}{5} \\ &> 3 & M_{12} &= \frac{1}{6} \end{aligned}$$

(Note: For a rectangular armature structure with the stops placed at its edges $M_{12} = \frac{1}{4}$, $P_{12} = \frac{1}{2}$ and force ratios in the neighborhood of 8 are desirable.)

X. ACKNOWLEDGMENT

The analytical treatment presented in this paper contains contributions by E. L. Norton, R. L. Peek, Jr., and the writer.

APPENDIX I

DERIVATION OF BASIC EQUATIONS OF MOTION THREE-DEGREE-OF-FREEDOM SYSTEM

(1) *Free Interval*

The motion of the armature will be described by the displacement at the stop points, x_1 , x_2 , x_3 . Let m be the mass and R the radius of gyration of the armature about the center of gravity. The latter is located by the dimensions $\ell_1 R$, $\ell_2 R$, and $\ell_3 R$ relative to the stop points (Fig. 3).

The rotation and displacement of the center of gravity is then

$$\left. \begin{aligned} x_h &= (x_2 - x_1) \frac{l_3}{l_1 + l_2} + x_3 \\ x_v &= x_1 + (x_2 - x_1) \frac{l_1}{l_1 + l_2} \\ \theta &= \frac{x_2 - x_1}{R(l_1 + l_2)} \end{aligned} \right\} \quad (a)$$

From this the kinetic energy may be computed

$$\left. \begin{aligned} T &= \frac{1}{2}m(\dot{x}_h^2 + \dot{x}_v^2) + \frac{1}{2}mR^2\dot{\theta}^2 \\ &= \frac{\dot{x}_1^2(1 + l_2^2 + l_3^2) + \dot{x}_2^2(l_1^2 + l_3^2 + 1) + \dot{x}_3^2(l_1 + l_2)^2}{2(l_1 + l_2)^2} \\ &\quad + \frac{2\dot{x}_1\dot{x}_2(l_1l_2 - l_3^2 - 1) - 2\dot{x}_3\dot{l}_3(l_1 + l_2)(\dot{x}_1 - \dot{x}_2)}{2(l_1 + l_2)^2} \end{aligned} \right\} \quad (b)$$

Applying LaGrange's Equation to the above, the equations of motion are obtained:

$$\left. \begin{aligned} \frac{d}{dt} \left(\frac{\partial T}{\partial \dot{q}_r} \right) - \frac{\partial T}{\partial q_r} &= P_r \\ \frac{F_1}{m} &= \frac{\ddot{x}_1[l_2^2 + l_3^2 + 1] + \ddot{x}_2[l_1l_2 - l_3^2 - 1] - \ddot{x}_3l_3(l_1 + l_2)}{(l_1 + l_2)^2} \\ \frac{F_2}{m} &= \frac{\ddot{x}_1[l_1l_2 - l_3^2 - 1] + \ddot{x}_2[l_1^2 + l_3^2 + 1] + \ddot{x}_3l_3(l_1 + l_2)}{(l_1 + l_2)^2} \\ \frac{F_3}{m} &= \frac{-\ddot{x}_1l_3(l_1 + l_2) + \ddot{x}_2l_3(l_1 + l_2) + \ddot{x}_3(l_1 + l_2)^2}{(l_1 + l_2)^2} \end{aligned} \right\} \quad (c)$$

The Equations (3) may be solved for \ddot{x}_1 , \ddot{x}_2 , \ddot{x}_3 and the results integrated, yielding

$$\left. \begin{aligned} x_1 &= \frac{1}{2m} [C_{11}F_1 + C_{12}F_2 + C_{13}F_3]t^2 + \dot{x}_{10}t + x_{10} \\ x_2 &= \frac{1}{2m} [C_{21}F_1 + C_{22}F_2 + C_{23}F_3]t^2 + \dot{x}_{20}t + x_{20} \\ x_3 &= \frac{1}{2m} [C_{31}F_1 + C_{32}F_2 + C_{33}F_3]t^2 + \dot{x}_{30}t + x_{30} \end{aligned} \right\} \quad (d)$$

where

$$\left. \begin{aligned} C_{11} &= (e_1^2 + 1) & C_{13} &= C_{31} = e_1 e_3 \\ C_{22} &= (e_2^2 + 1) & C_{12} &= C_{21} = (1 - e_1 e_2) \\ C_{33} &= (e_3^2 + 1) & C_{23} &= C_{32} = -e_2 e_3 \end{aligned} \right\} \quad (3)$$

\dot{x}_{10} , \dot{x}_{20} , \dot{x}_{30} are the initial velocities, x_{10} , x_{20} , x_{30} the initial displacements for the free interval in question. Interpretation of the analytic results is simplified by the introduction of normalization. Let \dot{x}_a be \dot{x}_1 just before the "zero" impact and define

$$\left. \begin{aligned} y_i &= \frac{x_i}{\dot{x}_a \tau} = \frac{F_1}{\dot{x}_a^2 m} x_i & \dot{y}_i &= \frac{d}{d\left(\frac{t}{\tau}\right)} y_i = \frac{\dot{x}_i}{\dot{x}_a} \\ \tau &= \frac{\dot{x}_a m}{F_1} \end{aligned} \right\} \quad (2)$$

Dividing Equations (d) by $\dot{x}_a \tau$ yields the normalized equations of motion:

$$\left. \begin{aligned} y_1 &= \frac{1}{2} \left[C_{11} + C_{12} \frac{(F_2)}{(F_1)} + C_{13} \frac{(F_3)}{(F_1)} \right] \left(\frac{t}{\tau} \right)^2 + \dot{y}_{10} \left(\frac{t}{\tau} \right) + y_{10} \\ y_2 &= \frac{1}{2} \left[C_{21} + C_{22} \frac{(F_2)}{(F_1)} + C_{23} \frac{(F_3)}{(F_1)} \right] \left(\frac{t}{\tau} \right)^2 + \dot{y}_{20} \left(\frac{t}{\tau} \right) + y_{20} \\ y_3 &= \frac{1}{2} \left[C_{31} + C_{32} \frac{(F_2)}{(F_1)} + C_{33} \frac{(F_3)}{(F_1)} \right] \left(\frac{t}{\tau} \right)^2 + \dot{y}_{30} \left(\frac{t}{\tau} \right) + y_{30} \end{aligned} \right\} \quad (1)$$

(2) Impact Interval

The change of velocity at point "i" due to an impact at "i" is, by definition of the coefficient of restitution "k":

$$\Delta \dot{x}_i = -(1 + k_i) \dot{x}_i \quad (e)$$

Since this velocity change occurs as rotation about the conjugate point as an instant center of rotation, the impact relationships may be written, for an impact at point "1",

$$\begin{aligned} \Delta \dot{x}_1 &= -(1 + k_1) \dot{x}_1 \\ \Delta \dot{x}_2 &= -(1 + k_1) \dot{x}_1 \frac{\left(\frac{R}{l_1} - l_2 R \right)}{\left(l_1 R + \frac{R}{l_1} \right)} \end{aligned}$$

$$\begin{aligned}
 &= (1 + k_1)\dot{x}_1 \frac{(\ell_1 \ell_2 - 1)}{(\ell_1^2 + 1)} = -\frac{C_{12}}{C_{11}} (1 + k_1)\dot{x}_1 \\
 \Delta \dot{x}_3 &= -(1 - k_1)\dot{x}_1 \frac{\ell_3 R}{\ell_1 R + \frac{R}{\ell_1}} \\
 &= -(1 + k_1)\dot{x}_1 \frac{\ell_1 \ell_3}{\ell_1^2 + 1} = -\frac{C_{13}}{C_{11}} (1 + k_1)\dot{x}_1
 \end{aligned}$$

Similarly it can be shown that impacts at points (2) and (3) follow the same pattern. The general impact relations for impact at point “*i*” are then

$$\dot{y}_{j0n} = \dot{y}_{je(n-1)} + K_{ji}\dot{y}_{ie(n-1)} \quad (6)$$

The first subscript indicates the coordinate, and the second subscript indicates the beginning (0) or end (*e*) of the free interval denoted by the third subscript.

The impact transfer coefficient K_{ji} relating a velocity change at point “*j*” to an impact at point “*i*”:

$$K_{ji} = -\frac{C_{ji}}{C_{ii}} (1 + k_i) \quad (7)$$

APPENDIX II

ANALYSIS OF REBOUND PATTERNS—ONE-DEGREE-OF-FREEDOM SYSTEM

The equation of motion of this system is

$$y_{1n} = \frac{1}{2}Ct'^2 + \dot{y}_{10n}t' + y_{10n} \quad (f)$$

where

$$C = C_{11} - \frac{C_{12}^2}{C_{22}} \quad (9)$$

$$t' = \frac{t}{\tau}$$

and is measured from the start of the particular interval of free motion in question. The impact relationship is

$$\dot{y}_{10n} = -k_1\dot{y}_{1e(n-1)}$$

The motion consists of a series of parabolic arcs having periods of $2\dot{y}_{10}/C$ in general, or $2/C$, $2k/C$, $2k^2/C$, \dots , $2k^{n-1}/C$. The time elapsed

is a convergent series and approaches, for a complete series:

$$\lim_{n \rightarrow \infty} \frac{2}{C} [1 + k + k^2 + \cdots k^n] = \frac{2}{C(1 - k)} \quad (10)$$

The maximum rebound amplitude in any interval is $-\dot{y}_{10n}/2C$. The maximum excursion occurs during the first bounce at $t' = 1/C$ and equals $-k^2/2C$.

APPENDIX III

ANALYSIS OF REBOUND PATTERNS—TWO-DEGREE-OF-FREEDOM SYSTEM

The equations of motion of this system are

$$\left. \begin{aligned} y_1 &= \frac{1}{2}At'^2 + \dot{y}_{10n}t' + y_{10n} \\ y_2 &= \frac{1}{2}Bt'^2 + \dot{y}_{20n}t' + y_{20n} \end{aligned} \right\} \quad (g)$$

$$\left. \begin{aligned} \text{where } A &= C_{11} + C_{12}f \\ B &= C_{12} + C_{22}f \\ f &= \frac{F_2}{F_1} \\ t' &= \frac{t}{\tau} \text{ measured from the start of the par-} \\ &\quad \text{ticular free interval in question.} \end{aligned} \right\} \quad (h)$$

A. Complete Front Series

At the beginning of a front series

$$\left. \begin{aligned} y_1 &= 0 \\ \dot{y}_1 &= \dot{y}_{1e0} \\ y_2 &= y_{2e0} \\ \dot{y}_2 &= \dot{y}_{2e0} \end{aligned} \right\} \quad (i)$$

In a manner analogous to that for the one-degree-of-freedom system each front impact reduces \dot{y}_1 to $-k_1\dot{y}_1$. Therefore, after the n^{th} impact,

$$\dot{y}_{10n} = -k_1^n \dot{y}_{1e0}$$

and the time elapsed in the n^{th} interval is

$$T_n = \frac{2k_1^n}{A} \dot{y}_{1e0} \quad (j)$$

At the heel, from (g), the heel velocity preceding the n^{th} impact is

$$\dot{y}_{2e(n-1)} = \dot{y}_{20(n-1)} + Bt' \quad (\text{k})$$

The velocity change during the $(n - 1)$ interval is then equal to BT_{n-1} . From Equations (6), (7) and (12), the change in velocity during the n^{th} impact is $-P_{12}(1 + k_1)k_1^{n-1}\dot{y}_{1e0}$.

The total change of \dot{y}_2 between impacts is then

$$\dot{y}_{20n} - \dot{y}_{20(n-1)} = BT_{n-1} - P_{12}(1 + k_1)k_1^{n-1}\dot{y}_{1e0}$$

Similarly in preceding intervals:

$$\begin{aligned} \dot{y}_{20(n-1)} - \dot{y}_{20(n-2)} &= BT_{n-2} - P_{12}(1 + k_1)k_1^{n-2}\dot{y}_{1e0} \\ &\vdots \\ \dot{y}_{202} - \dot{y}_{201} &= BT_1 - P_{12}(1 + k_1)k_1\dot{y}_{1e0} \\ \dot{y}_{201} - \dot{y}_{2e0} &= -P_{12}(1 + k_1)\dot{y}_{1e0} \end{aligned}$$

By addition of the above

$$\begin{aligned} \dot{y}_{20n} - \dot{y}_{2e0} &= B \sum_{m=1}^{n-1} T_m - P_{12}(1 + k_1) \sum_{m=0}^{n-1} k_1^m \dot{y}_{1e0} \\ &= \frac{2B}{A} \dot{y}_{1e0} \sum_{m=1}^{n-1} k_1^m - P_{12}(1 + k_1) \dot{y}_{1e0} \sum_{m=0}^{n-1} k_1^m \end{aligned}$$

The summations may be evaluated, yielding

$$\dot{y}_{20n} - \dot{y}_{2e0} = \left[\frac{2B}{A} \frac{k_1 - k_1^n}{1 - k_1} - P_{12}(1 + k_1) \frac{1 - k_1^n}{1 - k_1} \right] \dot{y}_{1e0} \quad (1)$$

To evaluate the displacements at the heel, Equation (g) yields

$$y_{20n} - y_{20(n-1)} = \dot{y}_{20(n-1)} T_{n-1} + \frac{1}{2} B T_{n-1}^2$$

Adding these expressions for intervals 0 to n ; the total change in y_2 is

$$\begin{aligned} y_{20n} - y_{201} &= \sum_{m=1}^{n-1} \dot{y}_{20m} T_m + \frac{1}{2} B \sum_{m=1}^{n-1} T_m^2 \\ &= \frac{2k_1(1 - k_1^{n-1})}{A(1 - k_1)} \dot{y}_{1e0} \dot{y}_{2e0} \\ &\quad + \left[\frac{2B(k_1^2 - 2k_1^{n+1} + k_1^{2n})}{A^2(1 - k_1)^2} \right. \\ &\quad \left. - \frac{2P_{12}k_1(1 - k_1^n - k_1^{n-1} + k_1^{2n-1})}{A(1 - k_1)^2} \right] \dot{y}_{1e0}^2 \end{aligned} \quad (\text{m})$$

Expressions for an initial series may be obtained by setting $\dot{y}_{1e0} = 1$, $\dot{y}_{2e0} = y_{2e0} = 0$, and, finally, for an initial complete series $m \rightarrow \infty$ and hence $k^m \rightarrow 0$, and Equations (l) and (m) become

$$\left. \begin{aligned} \dot{y}_{2e\infty} &= \frac{2k_1}{A(1+k_1)^2} \left[\frac{Bk_1}{A} - P_{12} \right] \\ \dot{y}_{2e\infty} &= \frac{1}{1-k_1} \left[\frac{2Bk_1}{A} - P_{12}(1+k_1) \right] \end{aligned} \right\} \quad (17)$$

B. Complete Heel Series

For heel series, Equations (l) and (m) may be used by interchanging the initial velocities, accelerations, and impact transfer coefficients for those relating to heel motion:

$$\dot{y}_{10n} - \dot{y}_{1e0} = \left[\frac{2A}{B} \frac{k_2 - k_2^n}{1 - k_2} - \frac{M_{12}(1+k_2)}{P_{12}} \frac{1 - k_2^n}{1 - k_2} \right] \dot{y}_{2e0} \quad (n)$$

$$\begin{aligned} y_{10n} - y_{1e0} &= \frac{2k_2(1 - k_2^{n-1})}{B(1 - k_2)} \dot{y}_{1e0} \dot{y}_{2e0} \\ &+ \left[\frac{2A(k_2^2 - 2k_2^{n+1} + k_2^{2n})}{B^2(1 - k_2^2)} - \frac{2M_{12}k_2(1 - k_2^n - k_2^{n-1} + k_2^{2n-1})}{BP_{12}(1 - k_2^2)} \right] \dot{y}_{2e0}^2 \end{aligned} \quad (o)$$

An initial heel series occurs when the heel strikes first after the "zero" impact. The first heel impact then occurs $T_1 = 2P_{12}/B(1+k_1)$ after the zero impact and the initial conditions are

$$\dot{y}_{2e1} = P_{12}(1+k_1)$$

$$\dot{y}_{1e1} = -k_1 + AT_1 = \frac{2AP_{12}}{B}(1+k_1) - k_1$$

$$y_{1e1} = -k_1T_1 + \frac{1}{2}AT_1^2 = \frac{2P_{12}}{B}(1+k_1) \left[\frac{AP_{12}}{B}(1+k_1) - k_1 \right]$$

Substitution of the above into (n) yields

$$\dot{y}_{10n} = -k_1 + \frac{1+k_1}{1-k_2} \left[\frac{2AP_{12}}{B}(1-k_2)^n - M_{12}(1+k_2)(1-k_2^n) \right] \quad (p)$$

The corresponding expression for y_{10n} is quite involved. For the special case of $k = k_1 = k_2$

$$y_{10n} = \frac{AP_{12}(1+k)^2}{B} \left[\frac{AP_{12}}{B} \left(\frac{1-k^n}{1-k} \right)^2 = \frac{k(1-k^n)}{1-k^2} - \frac{M_{12}(1-k^n)(k-k^n)}{(1-k)^2} \right] \quad (q)$$

If the initial series is a complete series, $n \rightarrow \infty$ and

$$\left. \begin{aligned} y_{10\infty} &= \frac{AP_{12}(1+k)^2}{B(1-k)^2} \left[\frac{AP_{12}}{B} - \frac{k(1-k)}{1+k} - M_{12}k \right] \\ \dot{y}_{1e\infty} &= \frac{1+k}{1-k} \left[\frac{2AP_{12}}{B} - \frac{k(1-k)}{(1+k)} - M_{12}(1+k) \right] \end{aligned} \right\} \quad (22)$$

C. Partial Front Series

The worst rebound occurs when heel and front impacts occur nearly simultaneously, with the front hitting first. From Equation (m) for an initial front series, this requires that

$$\frac{B}{AP_{12}} = Q = \frac{1-k^n}{k-k^n} \quad (30)$$

After n front impacts conditions are given by Equations (14) and (19) with $y_1 = y_2 = 0$, and

$$\frac{T+V}{T_0} = 1 - (1-M_{12})(1-k^{2n}) = k^{2n} + \frac{M_{12}\dot{y}_2^2}{P_{12}^2} - \frac{2M_{12}k^n\dot{y}_2}{P_{12}}$$

This may be solved for $\dot{y}_2 = P_{12}(1-k^n)$. The maximum front excursion now possible is that for a complete series of heel impacts. The above value of \dot{y}_2 in Equation (24) yields

$$2CY_1 = M_{12} + (1-M_{12})[k^{2n} - M_{12}(1-k^n)^2] \quad (31)$$

D. Partial Heel Series

The worst rebound occurs again when heel and front impacts occur nearly simultaneously, with the front hitting first. From Equation (9) for an initial heel series, this requires that

$$\frac{B}{AP_{12}} = Q = \frac{1-k^{n+1}}{\frac{k(1-k)}{1+k} + k(1-k^n)M_{12}} \quad (32)$$

After n heel impacts $\dot{y}_2 = P_{12}(1 + k)k^n$ and from Equations (19) and (23)

$$\begin{aligned}\frac{T + V}{T_0} &= 1 - (1 - M_{12})(1 - k^2) - M_{12}(1 - M_{12})(1 + k)^2(1 - k^{2n}) \\ &= \dot{y}_1^2 - 2M_{12}(1 + k)k^n\dot{y}_1 + M_{12}(1 + k)^2k^{2n}\end{aligned}$$

This may be solved for $\dot{y}_1 = k - M_{12}(1 + k)(1 - k^n)$ and after the front impact immediately following:

$$\dot{y}_2 = P_{12}(1 + k)k^n - P_{12}(1 + k)[k - M_{12}(1 + k)(1 - k^n)]$$

The maximum front excursion now possible is that for a complete series of heel impacts. The above value for \dot{y}_2 in Equation (24) yields

$$\begin{aligned}2CY_1 &= 1 - (1 - M_{12})(1 - k^2) \{1 + [k - M_{12}(1 + k)(1 - k^n)]^2\} \\ &\quad - M_{12}(1 - M_{12})(1 + k)^2 \{1 - k^{2n} \\ &\quad + [k - k^n - M_{12}(1 + k)(1 - k^n)]^2\}\end{aligned}\quad (33)$$

APPENDIX IV

SUMMARY OF NOTATION

$$A = C_{11} + C_{12} + C_{12}f$$

$$B = C_{12} = C_{22} + C_{22}f$$

$$C = C_{11} - \frac{C_{12}^2}{C_{22}}$$

$$C_{11} = 1 + l_1^2 \quad C_{12} = C_{21} = [1 - l_1 l_2]$$

$$C_{22} = 1 + l_2^2 \quad C_{13} = C_{31} = l_1 l_3$$

$$C_{33} = 1 + l_3^2 \quad C_{23} = C_{32} = l_2 l_3$$

$$f = \frac{F_2}{F_1}$$

F_1 = front tensioning force

F_2 = heel tensioning force

k_1 = coefficient of restitution at vertical front stop

k_2 = coefficient of restitution at vertical heel stop

$$K_{ji} = -\frac{C_{ji}}{C_{ii}}(1 + k_i)$$

l_1R = vertical front stop location relative to c.g.

l_2R = vertical heel stop location relative to c.g.

l_3R = horizontal heel stop location relative to c.g.

m = mass of armature

$$M_{ij} = \frac{C_{ij}^2}{C_{ii}C_{ij}}$$

$$P_{ij} = \frac{C_{ij}}{C_{ii}}$$

$$Q = \frac{1 + \frac{P_{12}}{M_{12}f}}{1 + P_{12}f} = \frac{t_1}{t_2} \frac{1 + k_1}{k_1}$$

R = radius of gyration of armature about center of gravity

$$\tau = \frac{\dot{x}_a m}{F_1}$$

t = time

$$t' = \frac{t}{\tau}$$

t_1 = basic period of front after "zero" impact

t_2 = basic period of heel after "zero" impact

T_n = duration of n^{th} free interval

x_1 = vertical front displacement

x_2 = vertical heel displacement

x_3 = horizontal heel displacement

\dot{x}_a = front velocity just prior to "zero" impact

$$y = \frac{x}{\dot{x}_a \tau}$$

Y_1 = greatest excursion (rebound) of front

Abstracts of Bell System Technical Papers Not Published in This Journal

A New Telephone Carrier System for Medium-Haul Circuits. R. S. CARUTHERS¹, H. R. HUNTLEY², W. E. KAHL¹, and L. PEDERSON¹. *Elec. Eng.*, **70**, pp. 692-693, Aug. 1951.

Low terminal costs and single-cable operation make this the first economically practical carrier system for medium-haul telephone circuits. Performance is not sacrificed for economy.

*A .15-Kw 500-Megacycle Grounded-Grid Tridode.** C. E. FAY¹, D. A. A. HALE¹, and R. J. KIRCHER¹. *Proc. I.R.E.*, **39**, pp. 800-803, July, 1951.

Short-Cut Method Aids Figuring Exhaust and Collecting Systems. W. H. FOGLE³. *Heating, Piping and Air Cond.*, **23**, pp. 75-78, July, 1951.

A simple, workable method of determining static losses in industrial exhaust and collecting systems is explained here by means of a sample problem and its step-by-step solution. The method might be described as an "averaging out" process, whereby all duct sizes and lengths are averaged out with the cfm and velocity to give a total linear footage of the average duct size and an average velocity.

*Arcing at Electrical Contacts on Closure. Part I. Dependence Upon Surface Conditions and Circuit Parameters.** L. H. GERMER¹. *Jl. Appl. Phys.* **22**, pp. 955-964, July, 1951.

In a low-voltage circuit the occurrence of an arc between approaching electrodes is dependent upon the nature of the surfaces and upon the circuit inductance. For carbon surfaces, or noble metal surfaces which have been "activated" by operation in various organic vapors resulting in a carbonaceous layer, the limiting circuit inductance is somewhat above 10^{-3} h, which is much higher than the limiting inductance for clean noble metal surfaces. This activation by organic vapors occurs for noble metals only and for certain vapors; for example, benzene derivatives. In the case of silver and benzene vapor, it has been shown that the activation is due to adsorption of benzene onto a greasy surface layer and its decomposition there by the heat of subsequent closures. A metal surface, which has been activated by organic vapor, remains active indefinitely if there

* A reprint of this article may be obtained on request.

¹ Bell Tel. Labs.

² A. T. & T. Co.

³ W. E. Co.

is no arcing at the surfaces; but with continued operation and accompanying arcing, the activating material is burned away, and the surface returns to the inactive condition if no activating vapor is supplied.

Arc voltages, which are independent of current and of ambient gas, as far as tested, have been measured for a number of metals and for carbon; the arc voltage for carbon is quite erratic in the range between 20 and 30 volts, but for each of a number of metals the arc voltage is steady.

Arcing at noble metal surfaces, similar to that induced by carbonaceous material from organic vapors, can be produced also by insulating particles or insulating films. The active condition gradually disappears with continued arcing, unless there is a steady supply of insulating material to the surface.

The minimum arc current has been measured to be 0.6 amp for active silver and for carbon, and 0.03 amp for inactive silver. These are the currents at which an established arc is extinguished.

*Iron-Silicon Alloys Heat Treated in a Magnetic Field.** M. GOERTZ¹. *Jl. Appl. Phys.*, **22**, pp. 964-965, July, 1951.

Heat treatment in a magnetic field has been found effective for iron-silicon alloys between two per cent and ten per cent silicon, the highest maximum permeability being obtained at about 6.5 per cent silicon. In a single crystal of this composition, magnetized parallel to a (100) direction, the hysteresis loop is squared by the magnetic anneal and the maximum permeability is increased from 50,000 to 3,800,000, the highest value yet reported.

Domain Boundary Motion in Ferroelectric Crystals and the Dielectric Constant at High Frequency. C. KITTEL¹. Letter to the Editor. *Phys. Rev.*, **83**, p. 458, July 15, 1951.

*A Method for Determining the Propagation Constants of Plastics at Ultrasonic Frequencies.** H. J. McSKIMIN¹. *J. Acoust. Soc. Am.*, **23**, pp. 429-434, July, 1951.

A pulse technique particularly suited to dissipative materials is described for measuring attenuation and phase-shift constants of plastics, using either transverse or longitudinal waves in the frequency range of 5-50 mc.

A thin wafer of the material under test is placed between two identical fused silica buffers; and waves generated by quartz crystals at the ends of the assembly are transmitted simultaneously through the specimen in both directions. Comparison of transmitted and reflected components by means of a special balancing circuit provides information from which the complex propagation constant can be calculated, and hence dynamic rigidities and viscosities.

Illustrative data for polyethylene and Nylon are given.

* A reprint of this article may be obtained on request.

¹ Bell Tel. Labs.

Some Antecedents of Quality Control. E. C. MOLINA⁴. *Ind. Quality Control*, **8**, pp. 10-11, July, 1951.

Traveling-Wave Amplifier Measurements. F. E. RADCLIFFE¹. *Electronics*, **24**, pp. 110-111, Aug., 1951.

Rapid sweep-frequency technique used at 4,000 mc can be applied to all broad-band amplifier measurements. Oscilloscope display shows transmission accurate to 0.1 db and return-loss values up to 40 db.

*Kirchhoff's Formula, its Vector Analogue, and Other Field Equivalence Theorems.** S. A. SCHELKUNOFF¹. *Communications on Pure and Applied Math.*, **4**, pp. 43-59, June, 1951.

*Remarks Concerning Wave Propagation in Stratified Media.** S. A. SCHELKUNOFF¹. *Communications on Pure and Applied Math.*, **4**, pp. 117-128, June, 1951.

*An Achromatic Doublet of Silicon and Germanium.** R. G. TREUTING¹. *J. Opt. Soc. Am.*, **41**, pp. 454-456, July, 1951.

The semi-metals germanium and silicon have high transparency and high refractive indices over a wide range of infrared wavelengths and are stable to normal atmospheres. Their relative indices and dispersions make achromatic combinations possible; and designs are given for axially corrected doublets of relative apertures f:2 and f:1. The optical homogeneity of the materials is discussed: compositional variations are not considered an optical hazard, but there is evidence of structural imperfections in some specimens whose effect on optical properties remains to be evaluated.

*On the Motion of Gaseous Ions in a Strong Electric Field. I.** G. H. WANNIER¹. *Phys. Rev.*, **83**, pp. 281-289, July 15, 1951.

This paper applies the Boltzmann method of gaseous kinetics to the problem of positive ions moving through a gas under the influence of a static, uniform electric field. The ion density is assumed to be vanishingly low, but the field is taken to be strong; that is, the energy which it imparts to the ions is not assumed negligible in comparison to thermal energy. Attention is focused upon the computation of velocity averages, and the drift velocity in particular, rather than a complete knowledge of the entire velocity distribution. It is shown in Sections C and E that the problem so formulated is completely soluble if the mean free time between collisions of ions and molecules is a constant; this is the case for the so-called polarization force between ions and molecules which predominates over other forces at low temperature. A method for obtaining averages to any desired accuracy in the general case is developed in Section D. The method is applied to the hard sphere model for the high field range and

* A reprint of this article may be obtained on request.

¹ Bell Tel. Labs.

⁴ Bell Tel. Labs., Retired.

mass ratio 1. An application of the resulting formula (43) to experimental material has been published earlier.

Evidence for the Noncubic High Temperature Phase of BaTiO₃. E. A. WOOD¹. Letter to the Editor. *J. Chem. Phys.*, **19**, p. 976, July, 1951.

*Seven-League Oscillator.** F. B. ANDERSON¹. *Proc. I.R.E.*, **39**, pp. 881-890, Aug., 1951.

A bridge-type RC oscillator is described which is continuously adjustable over a frequency range of 20 cps to 3 mc in one sweep of a two-gang linear potentiometer control. Tracking requirements of the two-gang control are not severe. The output is available in four phases, and the frequency is an approximately logarithmic function of the linear potentiometer setting. Practical limits of the frequency range are tentatively 0.01 cps and 10 mc. Accuracy of setting of the order of one per cent is attainable with ordinary components. Frequency stability is of the order of 2 per cent per db of tube gain variation.

*Semi-Conductor Surface Phenomena.** W. H. BRATTAIN¹. *Semi-Conducting Materials*, H. K. HENISCH, ed., pp. 37-46. Proceedings of a conference held at the University of Reading (July 10-15, 1950) London, Butterworths, 1951.

Developments in the understanding and interpretation of phenomena occurring at the surface of a semi-conductor are reviewed. The development starts with the Mott-Schottky theory of the space charge layer. Bardeen's concept of a space charge layer due to 'surface states' explained the independence of rectification on contact potential and Meyerhoff's small values for the contact potential between n- and p-type silicon. Shockley and Brattain observed that this contact potential difference increased with impurity concentration. Illumination of a silicon surface produced hole and electron pairs in the space charge layer. The potential of the surface changed until the photocurrent was balanced by a conduction current. The relation between photo-current and potential change was of the same form as a forward characteristic for a rectifying contact. In an experiment similar to Becquerel's, using water as an electrolyte, the surface may be biased in the reverse direction. When so biased the response to modulated light gives the differential resistance of the space charge layer. This resistance increases rapidly with reverse bias and the time constant of the layer increases, both agreeing qualitatively with theory. This experimental method of measuring changes in surface potential caused by illumination permits determination of the properties of the space charge layer at the free surface of a semi-conductor.

*The Calculation of Traveling-Wave-Tube Gain.** C. C. CUTLER¹. *Proc. I.R.E.*, **39**, pp. 914-917, Aug., 1951.

* A reprint of this article may be obtained on request.

¹ Bell Tel. Labs.

Essential information for calculation of traveling-wave-tube gain is summarized and condensed in this paper. The important relations are documented, presented in a concise form for simplified computation, and developed as a nomograph. The conclusions have been found to be in agreement with measurements on six different tube designs.

*Thermal Variation of Young's Modulus in some Fe-Ni-Mo Alloys.** M. E. FINE¹ and W. C. ELLIS¹. References. *Jl. Metals*, **3**, pp. 761-764, Sept., 1951.

*A Broad-Band Transcontinental Radio Relay System.** T. J. GRIESER¹ and A. C. PETERSON¹. *Elec. Eng.*, **70**, pp. 810-815, Sept., 1951.

Spanning the continent from coast to coast, this microwave relay system provides six channels, each of which can carry one television circuit or hundreds of telephone circuits. Some features of this vast network are described.

*An Improved Telephone Set.** A. H. INGLIS¹ and W. L. TUFFNELL¹. *Elec. Eng.*, **70**, pp. 770-775, Sept., 1951.

The familiar telephone set has undergone numerous changes which will provide better service at lower cost than do present models. Increased transmitting and receiving gain, better sidetone control, broader frequency response, faster dialing, simple ringing control, and a trim appearance are some of the features of the new design.

The Institutes for Basic Research—Their Contribution to National Strength. M. J. KELLY¹, pp. 11-23. *Applied Research is Not Enough*, (booklet). Addresses at the Dedication of the Institutes for Basic Research, The University of Chicago, May 16, 1951.

The Crystal Clock. W. A. MARRISON¹. *Science Marches On*, JAMES STOKLEY, ed., N. Y., Ives Washburn, Inc., 1951, pp. 303-309.

Observations of Zener Current in Germanium p-n Junctions. K. B. McAFEE¹, E. J. RYDER¹, W. SHOCKLEY¹, and M. SPARKS¹. Letter to the Editor. *Phys. Rev.*, **83**, pp. 650-651, Aug. 1, 1951.

*Experimental Radio-Telephone Service for Train Passengers.** N. MONK¹. *Proc. I.R.E.*, **39**, pp. 873-881, Aug., 1951.

Experimental public radio-telephone service for train passengers was inaugurated by the Bell Telephone System several years ago. Initial installations operated in conjunction with a series of urban mobile base stations. More recently, highway mobile systems have been used for this service, and this paper describes a typical train installation operating through a highway channel. All

* A reprint of this article may be obtained on request.

¹ Bell Tel. Labs.

of these early systems utilized an attendant on the train. The cost of providing an attendant has, in some cases, been found excessive. Consequently, experiments have been initiated in which a coin box is used on the train. The arrangements for this purpose are also described.

*The Magneto-Resistance Effect in Oriented Single Crystals of Germanium.** G. L. PEARSON¹ and H. SUHL¹. *Phys. Rev.*, **83**, pp. 768-776, Aug. 15, 1951.

This paper describes an extensive study of the magneto-resistance effect in germanium as a function of crystal orientation. Experimental measurements establish the constants involved in the dependence of the effect on orientation of magnetic field and electric relative to the crystal axes. The measurements are internally consistent with existing phenomenological theory based on cubic crystal symmetry, in which terms involving the magnetic field to higher than the second order are neglected. It is shown that such deviations as do occur arise from higher terms in the field, since an extension of the phenomenological theory to the fourth order predicts their symmetry. Relations are established between the experimentally observed phenomenological constants and those constants appearing in existing magneto-resistance electronic theories. It is concluded that no electronic theory yet worked out is entirely consistent with experiment. The present electronic theories are special cases of a very general theory recently proposed by Shockley, and it is possible that agreement can be obtained as soon as the computational difficulties of the latter theory are overcome.

New Phenomena of Electronic Conduction in Semi-Conductors. W. SHOCKLEY¹. *Semi-Conducting Materials*, H. K. HENISCH, ed., pp. 26-36. Proceedings of a conference held at the University of Reading (July 10-15, 1950) London, Butterworths, 1951.

The semi-conductors silicon and germanium may be discussed as insulators the electronic structure of which is disturbed. Excess electrons, which act as negatively charged current carriers, may be present as may be 'holes' or places where electrons are missing from the valence-bond structure. Holes act as positively charged current carriers. In ordinary electronic conduction the flow of current carriers is substantially incompressible so that the density of carriers remains constant. When a new transistor phenomenon known as 'carrier injection' occurs, however, the total density of holes and electrons may be greatly increased and this modulation of the electronic structure may be used both for scientific investigation and for practical amplification. In particular, carriers may be injected at a predetermined time and place into a known uniform electric field and their transit time to another place accurately timed by detecting their arrival with a "collector" point. Drift velocities and mobilities may be measured precisely in this way with a directness unattainable by pre-transistor

* A reprint of this article may be obtained on request.

¹ Bell Tel. Labs.

methods involving conductivity and Hall effect. These new experiments and their related theories may furnish a foundation for a new engineering science of transistor electronics.

*Domain Patterns on Nickel.** H. J. WILLIAMS¹ and J. G. WALKER¹. *Phys. Rev.*, **83**, pp. 634-636, Aug. 1, 1951.

Domain patterns have been observed on two single crystals of nickel cut in the form of hollow parallelograms. The length of the sides were parallel to the (111) directions in one specimen and to the (110) directions in the other. The crystals show domain structures with the three types of domain boundaries which are to be expected from a material having the directions of easy magnetization along the (111) directions. Domain boundary movement under the influence of an applied magnetic field was observed.

*Polymorphism in Potassium Niobate, Sodium Niobate, and Other ABO₃ Compounds.** E. A. WOOD¹. Bibliography. *Acta Cryst.*, **4**, pp. 353-362, July, 1951.

The first part of this paper presents the results of optical and X-ray studies of the perovskite-type crystals, potassium niobate and sodium niobate. Potassium niobate is orthorhombic at room temperature, changing to tetragonal at about 225°C. and cubic near 435°C. Sodium niobate is orthorhombic at room temperature, changing to tetragonal at about 370°C. and to cubic at about 640°C.

The second part of the paper discusses relations among the structures of the ABO₃ compounds.

*Subjective Sharpness of Additive Color Pictures.** M. W. BALDWIN¹. *Proc. I.R.E.*, **39**, pp. 1173-1176, Oct., 1951.

This is a report on the first numerical results to come from a laboratory experiment on the subjective sharpness of additive three-color pictures. The sharpness factor is isolated by using out-of-focus projection (of slides) instead of actual television transmission.

An observer's acuity for defocus is greatest for the green component and least for the blue component, in an additive three-color picture. When the same picture is reproduced in monochrome (white, red, green, or blue) at the same brightness, the observer's acuity for defocus is equal to that found for the green component.

*Frequency-Modulation Terminal Equipment for the Transcontinental Relay System.** J. G. CHAFFEE¹ and J. B. MAGGIO¹. *Elec. Eng.*, **70**, pp. 880-883, Oct., 1951.

To meet the exacting requirements of the new transcontinental microwave relay system, specially designed frequency-modulation terminal equipment was

* A reprint of this article may be obtained on request.

¹ Bell Tel. Labs.

constructed. The terminal transmitter converts either message or television signals to a frequency-modulated signal centered on 70 mc and the terminal frequency-modulation receiver recovers these signals, thus providing a link between the relay system and other telephone facilities.

*Observer Reaction to Low-Frequency Interference in Television Pictures.**
A. D. FOWLER¹. *Proc. I.R.E.*, **39**, pp. 1332-1336, Oct., 1951.

This paper presents results of tests to determine how much low-frequency interference can be tolerated in black-and-white television pictures. Various levels of single low-frequency interference were superimposed on a locally transmitted television picture. Observers viewed the picture and rate the disturbing effect of each level of the interference. Ratings were made in terms of preworded comments ranging from "not perceptible" to "unusable." Interfering frequencies from 48 to 90 cycles per second were employed.

Just visible interference appears as a flicker. The rate of flicker is the difference between interfering and 60-cycle field frequencies. The most disturbing interference produced a flicker rate of 5 or 6 cycles per second. To be tolerated, peak-to-peak amplitude of this interference had to be 54 db weaker than the peak-to-peak amplitude of the television signal (including synchronizing pulse). For flicker rates of 0.5 and 12 cycles per second, the amount of interference which could be tolerated was larger by 14 and 3 db, respectively.

*Arcing at Electrical Contacts on Closure. Part II. The Initiation of an Arc.** L. H. GERMER¹. *Jl. Appl. Phys.*, **22**, pp. 1133-1139, Sept., 1951.

The capacity of the plates of an oscilloscope charged to 35 or 40 volts is discharged repeatedly by approaching electrodes of carbon, active silver, and inactive silver. Facts about the discharges, which are arcs of very short duration, are inferred from resulting open circuit potentials and calculated electrode separations.

The separation at the first arc varies in different experiments but corresponds on the average to a nominal electric field of 0.6×10^6 volts/cm for carbon or active silver and to 2×10^6 volts/cm for inactive silver. Each arc is initiated by a very small number of field emission electrons. The hypothesis that a single electron may perhaps be sufficient is consistent with observations at later stages of each closure when the electrodes are closer and the field much higher.

The earlier observation, that the potential across a short arc is constant and independent of current, is not true if the arc time is sufficiently short. For active silver a time comparable with 2×10^{-8} sec is required to establish the steady arc voltage characteristic of later stages of arcs which last longer than this. The initial time during which the potential is decreasing toward its final steady value is 100 times the transit time of a silver ion across the gap.

* A reprint of this article may be obtained on request.

¹ Bell Tel. Labs.

Computation of Control Limits for p-Charts When the Samples Vary in Size. H. L. JONES⁵. *Ind. Quality Control*, **8**, pp. 26-27, Sept., 1951.

The Design of Switching Circuits. W. KEISTER¹, A. E. RITCHIE¹, and S. H. WASHBURN¹. N. Y., Van Nostrand, 1951. 556 pp. (Bell Telephone Laboratories Series).

This is the first published textbook in its field. It presents, first, the fundamental design principles of switching circuits composed of discrete-valued switching elements. Most of the discussion concerns two-valued elements, with greatest emphasis placed on electromagnetic relays. Chapters cover basic circuit paths, the logical interpretation of requirements, and the techniques of constructing networks to fulfill these requirements. The symbolic methods of Boolean algebra and its application to the design of combinational and sequential circuits is covered. Later chapters cover various unifunctional circuits such as selecting, connecting, translating, counting, and lockout. Final chapters discuss methods of synthesising unifunctional circuit building blocks into larger circuits and systems.

Measurement of the Elastic Constants of Silicon Single Crystals and Their Thermal Coefficients. H. J. McSKIMIN¹, W. L. BOND¹, E. BUEHLER¹, and G. K. TEAL¹. Letter to the Editor. *Phys. Rev.*, **83**, p. 1080, Sept. 1, 1951.

Interest in the properties of silicon single crystals arising from their use as semiconductors has led us to make measurements of the elastic constants of two single crystals. Measurements of velocities of propagation for both shear and longitudinal waves were made in the crystals as described in a recent paper by McSkimin. Frequencies in the range 8-12 mc/sec were used.

The three independent elastic constants were evaluated, a density of 2.331 (measured by pycrometer) being used. Data and formulas used are summarized in Table I. Two crystals were measured—as indicated—with data obtained from the larger one being used to determine the elastic constants. Check measurements were made for the smaller crystal; and despite the less accurate "pulse overlap" technique used for two of the measurements, velocity agreement to within 0.15 per cent was obtained.

Both crystals were of a high degree of crystalline perfection as shown by etching and X-ray tests.

Domain Wall Relaxation in Nickel. W. P. MASON¹. Letter to the Editor. *Phys. Rev.*, **83**, pp. 683-684, Aug. 1, 1951.

Phase Transitions in Ferroelectrics. B. MATTHIAS¹. National Research Council, Comm. on Solids. *Phase Transformations in Solids*. Ed. by R. Smoluchowski, J. E. Mayer, W. A. Weyl. N. Y., Wiley, 1951. 660 pp.

¹ Bell Tel. Labs.

⁵ Ill. Bell Tel. Co

Under the name ferroelectrics are classified those materials which exhibit dielectric anomalies phenomenologically similar to the magnetic behavior of the ferromagnetics. Perhaps it would have been more logical to use the term Rochelle-electrics, thus emphasizing the similarity in the dielectric behavior to that of Rochelle salt.

In this paper the known ferroelectrics are listed first, and then there follows a discussion of the various theories which have been created to explain them.

*Data on Random-Noise Requirements for Theater Television.** P. MERTZ¹. *Jl. S.M.P.T.E.*, **57**, pp. 89-107, Aug., 1951.

Provisional evaluation of permissible random noise for theater television is considered from several sources of information. These cover broadcast television experience and the graininess in motion picture film; the requirements deduced from the various sources generally agree. For broadcast television, a frequency weighting and limit on weighted noise power have been used. The finer picture detail of theater television presumes a lower permissible random noise. Changes in weighting curve are discussed. A limit figure of noise is suggested, which is comparable to graininess effects in motion pictures, though slightly more severe than present published performance on camera tubes.

*A Spatial Harmonic Traveling-Wave Amplifier for Six Millimeters Wavelength.** S. MILLMAN¹. *Proc. I.R.E.*, **39**, pp. 1035-1043, Sept., 1951.

This paper describes a traveling-wave amplifier in which the electron beam interacts with a spatial harmonic of an electromagnetic wave propagating along an array of resonator slots. The result is a considerable reduction in operating beam voltage for a given physical separation of the circuit elements. This type of amplifier operating at about 1,200 volts has yielded net power gains of about 18 db in the 6-mm wavelength region. A magnetic field of about 1,600 gauss is sufficient for proper beam focusing. Aside from small variations of gain with frequency that is caused by internal reflections, the bandwidth is of the order of 3 per cent.

*Form of Transient Currents in Townsend Discharges with Metastables.** J. P. MOLNAR¹. *Phys. Rev.*, **83**, pp. 933-940, Sept. 1, 1951.

The form of the current is calculated for a Townsend discharge stimulated by a pulsed light beam, with particular reference to the current component initiated by metastable effects. The calculation is directed particularly to the development of methods for quantitative interpretation of current patterns observed experimentally.

*Studies of γ -Processes of Electron Emission Employing Pulsed Townsend Discharges on a Millisecond Time Scale.** J. P. MOLNAR¹. *Phys. Rev.*, **83**, pp. 940-952, Sept. 1, 1951.

* A reprint of this article may be obtained on request.

¹ Bell Tel. Labs.

The relative amounts of electron emission from the cathode in a Townsend discharge caused by ions, photons, and metastables have been studied experimentally for several cathodes in argon, using pulsed-light stimulation of the discharge. The current initiated by metastables exhibits a slow build-up and decay, thus permitting easy separation from the faster rising effects of gas ionization and electron emission by photons and ions. Time constant studies of the slow component yielded a diffusion constant for metastable argon atoms of $45 \text{ cm}^2 \text{ sec}^{-1}$ at one millimeter pressure. The efficiencies of electron emission by metastables and ions was found to be closely the same, while the quantum yield for photon emission was found to be generally smaller.

*Electrical Properties of $\alpha\text{Fe}_2\text{O}_3$ and $\alpha\text{Fe}_2\text{O}_3$ Containing Titanium.** F. J. MORIN¹. *Phys. Rev.*, **83**, pp. 1005–1010, Sept. 1, 1951.

Electrical conductivity, Hall effect, and Seebeck effect have been measured on two sets of polycrystalline samples of $\alpha\text{Fe}_2\text{O}_3$ and $\alpha\text{Fe}_2\text{O}_3$ containing from 0.05 to 1.0 atomic per cent titanium (n-type impurity). One set of samples contained 0.6 atomic per cent excess of iron (n-type impurity), the second set contained 0.6 atomic per cent deficiency of iron (p-type impurity).

The conductivity of pure $\alpha\text{Fe}_2\text{O}_3$ is independent of this amount of stoichiometric deviation. The slope of the log conductivity vs reciprocal temperature plot is 1.17 eV and the intercept at $1/T = 0$ is $2.1 \times 10^4 \text{ ohm}^{-1} \text{ cm}^{-1}$. Room temperature conductivity varies from $-10^{-14} \text{ ohm}^{-1} \text{ cm}^{-1}$ (extrapolated) for pure $\alpha\text{Fe}_2\text{O}_3$ to $0.3 \text{ ohm}^{-1} \text{ cm}^{-1}$ for $\alpha\text{Fe}_2\text{O}_3$ containing 1.0 atomic per cent titanium.

The measured Hall voltages seem to result entirely from magnetization of the samples, which are weakly ferromagnetic, and disappear above the ferromagnetic Curie temperature.

The temperature variations of the Fermi level are determined from Seebeck data. The temperature variations of carrier concentration are determined from Fermi level and of mobility from carrier concentration and conductivity for some samples. Carrier concentration results indicate that each added titanium ion donates approximately one electron to the conduction process. Mobilities are found to be less than $2.0 \text{ cm}^2/\text{volt sec}$, suggesting that conduction involves electrons in the d level of iron.

*Acceptance Inspection of Purchased Material.** J. E. PALMER² and E. G. D. PATERSON¹. *Ind. Quality Control*, **8**, pp. 15–19, Sept., 1951.

This paper describes some of the principles and procedures employed in the inspection of purchased material in the form of components or finished products. The authors' experience has been largely with procedures used in the Bell System, and the illustrations have therefore been drawn from this source. It is felt, however, that considerations leading to the choice of specific inspection tech-

* A reprint of this article may be obtained on request.

¹ Bell Tel. Labs.

² W. E. Co.

niques will be generally applicable even though the number and volume of items purchased and the number of suppliers involved may in some cases differ widely. In this presentation, stress has been placed on a discussion of the broader gauge factors underlying the engineering planning of inspection procedures rather than on specific sampling and control techniques.

*Analysis of Audio-Frequency Atmospherics.** R. K. POTTER¹. *Proc. I.R.E.*, **39**, pp. 1067-1069, Sept., 1951.

Sound portrayal techniques used in studies of speech and noise reveal the structure of atmospheric disturbances well known to long-wave radio and ocean-cable engineers as "whistlers," "swishes," and "tweaks." It is suggested that renewed investigation of these effects, using modern analyzing tools, might yield information of considerable scientific interest.

*Reflection of Electromagnetic Waves from Slightly Rough Surfaces.** S. O. RICE¹. *Communications on Pure and Applied Math.*, **4**, pp. 351-378, Aug., 1951.

*Color Television and Colorimetry.** W. T. WINTRINGHAM¹. *Proc. I.R.E.*, **39**, pp. 1135-1172, Oct., 1951.

The high lights of the history of color measurement and of color photography are reviewed. Following this introduction, the principles of modern three-color colorimetry are developed from a hypothetical experiment in color matching. The conventional theory of "perfect color reproduction" by color television is built up from colorimetric background. Some of the difficulties to be expected in applying colorimetry to color television are brought out.

Finally, there is some discussion which tends to show that colorimetry may not be a sufficiently powerful tool to provide answers to all of the questions which will arise in the reproduction of scenes in color by television. The advantage of colorimetry as a background is indicated, however.

* A reprint of this article may be obtained upon request.

¹ Bell Tel. Labs.

Contributors to this Issue

CHARLES CLOS, C.E., New York University, 1927; New York Telephone Company, plant extension engineering, valuation and depreciation matters, intercompany settlements and tandem and toll fundamental plans, 1927-47. Pratt Institute, Evening School, Mathematics Instructor, 1946-49. Bell Telephone Laboratories, studies on development planning for local and toll switching systems and research in switching probability, 1947-. Member of A.I.E.E., New York Electrical Society, Mathematical Association of America, A.A.A.S., American Statistical Association, Iota Alpha, and Tau Beta Pi.

A. B. CRAWFORD, B.S. in E.E., Ohio State University, 1928; Bell Telephone Laboratories, 1928-. As a member of the Radio Research Department, he has been concerned with ultra short wave apparatus, measuring techniques, and propagation, and with microwave apparatus, measuring techniques, and propagation, as well as microwave radar and microwave antenna research. Member of I.R.E., Sigma Xi, Tau Beta Pi, Eta Kappa Nu, and Pi Mu Epsilon.

O. E. DE LANGE, B.S., University of Utah, 1930; M.A., Columbia University, 1937. Bell Telephone Laboratories, 1930-. Mr. De Lange has been engaged in radio research, including studies on high-frequency transmitters and receivers, frequency modulation, radar, broad-band systems, and pulse systems. Associate member of the I.R.E.

C. L. HOGAN, B.S. in Ch.E., Montana State College, 1942; M.S. in Physics, Lehigh University, 1947; Ph.D. in Physics, Lehigh, 1950. Anaconda Copper Mining Co., Great Falls, Montana, 1942-43. U. S. Navy, 1943-46. Instructor in Physics, Lehigh, 1947-50. Bell Telephone Laboratories, 1950-. Dr. Hogan has engaged in development work on boro-carbon resistors and microwave gyrators. Gold medal award for "Outstanding Engineer in Graduating Class," Montana State College, 1942. Letter of Merit from Chief of Naval Operations for work done in establishing and maintaining the acoustical torpedo shop at Pearl Harbor, 1944-46. Member of American Physical Society, Sigma Xi, Tau Beta Pi, and Phi Kappa Phi.

W. C. JAKES, JR., B.S.E.E., Northwestern University, 1944; M.S., Northwestern University, 1947; Ph.D., Northwestern University, 1949;

U.S. Navy, Airborne Radar Maintenance, 1944-46; Bell Telephone Laboratories, 1949-. Dr. Jakes has been engaged in microwave antenna and propagation studies. Member of I.R.E., Sigma Xi, Pi Mu Epsilon, and Eta Kappa Nu.

W. P. MASON, B.S. in E.E., University of Kansas, 1921; M.A., Ph.D., Columbia, 1928. Bell Telephone Laboratories, 1921-. Dr. Mason has been engaged principally in investigating the properties and applications of piezoelectric crystals, in the study of ultrasonics, and in mechanics. Fellow of the American Physical Society, Acoustical Society of America and Institute of Radio Engineers and member of Sigma Xi and Tau Beta Pi.

H. J. McSKIMIN, B.S., University of Illinois, 1937; M.S., New York University, 1940. Bell Telephone Laboratories, 1937-. Here he has worked chiefly on crystal filters, piezoelectric elements, ADP crystals, studies of the acoustic properties of liquids and solids. Fellow of Acoustical Society of America and Member of Eta Kappa Nu and Sigma Xi.

R. S. OHL, B.S. in Electro-Chemical Engineering, Pennsylvania State College, 1918; U. S. Army, 1918 (2nd Lieutenant, Signal Corps); Vacuum tube development, Westinghouse Lamp Company, 1919-21; Instructor in Physics, University of Colorado, 1921-22. Department of Development and Research, American Telephone and Telegraph Company, 1922-27; Bell Telephone Laboratories, 1927-. Mr. Ohl has been engaged in various exploratory phases of radio research, the results of which have led to numerous patents. For the past ten or more years he has been working on some of the problems encountered in the use of millimeter radio waves. Member of American Physical Society and Alpha Chi Sigma and Senior Member of the I.R.E.

E. E. SUMNER, B.M.E., Cooper Union, 1948, holding Schweinburg Scholarship throughout entire college curriculum; Instructor of Physics, Cooper Union, 1947-48; Non-resident instructor of Massachusetts Institute of Technology, *Probability and Statistics—Applications to Sampling and Quality Control*, summer, 1950; Bell Telephone Laboratories, 1948-. Mr. Sumner was given rotational assignments in apparatus, switching, and television transmission development and switching research, and has worked on a number of projects, including the card translator, the magnetic drum, the video transmission evaluator, and the vibrating reed selector. He is currently engaged in the development of wire-spring relay. Member of Tau Beta Pi and Pi Tau Sigma.

ROGER I. WILKINSON, B.S. in E.E., 1924, Professional Engineer (honorary), 1950, Iowa State College, 1924; Northwestern Bell Telephone Company, 1920-21; American Telephone and Telegraph Company, 1924-34; Bell Telephone Laboratories, 1934-43 and 1946-. U. S. War Department, Washington and South Pacific, 1943-45. Mr. Wilkinson has been engaged in applications of the mathematical theory of probability to telephone problems. Medal for Merit, 1946. Member of A.S.E.E.; A.S.A.; Institute of Mathematical Statistics; American Society for Quality Control; Associate Member of A.I.E.E.; and Member of Eta Kappa Nu; Tau Beta Pi; Phi Kappa Phi; and Pi Mu Epsilon.

T H E B E L L S Y S T E M

Technical Journal

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

VOLUME XXXI

MARCH 1952

NUMBER 2

Introduction to Formal Realizability Theory—I

BROCKWAY MC MILLAN 217

An Application of Boolean Algebra to Switching Circuit Design

ROBERT E. STAEHLER 280

Interaction of Polymers and Mechanical Waves

W. O. BAKER AND J. H. HEISS 306

The Reliability of Telephone Traffic Load Measurements by Switch
Counts

W. S. HAYWARD, JR. 357

Network Representation of Transcendental Impedance Functions

M. K. ZINN 378

Abstracts of Bell System Technical Papers Not Published in This
Journal

405

Contributors to This Issue

409

THE BELL SYSTEM TECHNICAL JOURNAL

PUBLISHED SIX TIMES A YEAR BY THE
AMERICAN TELEPHONE AND TELEGRAPH COMPANY

195 BROADWAY, NEW YORK 7, N.Y.

CLEO F. CRAIG, *President*

CARROLL O. BICKELHAUPT, *Secretary*

DONALD R. BELCHER, *Treasurer*

EDITORIAL BOARD

F. R. KAPPEL

O. E. BUCKLEY

H. S. OSBORNE

M. J. KELLY

J. J. PILLIOD

A. B. CLARK

R. BOWN

D. A. QUARLES

F. J. FEELY

PHILIP C. JONES, *Editor*

M. E. STRIEBY, *Managing Editor*

SUBSCRIPTIONS

Subscriptions are accepted at \$3.00 per year. Single copies are 75 cents each.

The foreign postage is 65 cents per year or 11 cents per copy.

PRINTED IN U.S.A.

Introduction to Formal Realizability Theory—I

By BROCKWAY McMILLAN

(Manuscript received October 15, 1951)

This paper offers a general approach to the realizability theory of networks with many accessible terminals. The methods developed are applied to give a complete characterization of all finite passive networks.

I. SUMMARY

1.0 A principal result of this paper is to characterize those matrices $Z(p)$, functions of the frequency parameter p , which can be realized as open-circuit impedance matrices of finite passive networks. This characterization is provided by the following theorem:

1.1 *Theorem*:* Let $Z(p)$ be an $n \times n$ matrix whose elements are $Z_{rs}(p)$, $1 \leq r, s \leq n$, where

- (i) Each $Z_{rs}(p)$ is a rational function
- (ii) $\overline{Z_{rs}(p)} = Z_{rs}(\bar{p})$ (the bar denotes complex conjugate)
- (iii) $Z_{rs}(p) = Z_{sr}(p)$
- (iv) For each set of real constants k_1, \dots, k_n , the function

$$\varphi_Z(p) = \sum_{r,s=1}^n Z_{rs}(p) k_r k_s$$

has a non-negative real part whenever $\operatorname{Re}(p) > 0$.

Then there exists a finite passive network, a $2n$ -pole, which has the impedance matrix $Z(p)$.

* Presented to the American Mathematical Society, April 17, 1948. Abstract 260, *Bulletin of the A.M.S.* No. 54, July, 1948.

Conversely, if a finite passive $2n$ -pole has an impedance matrix $Z(p)$, that matrix has the properties (i), (ii), (iii), (iv).

A formally identical dual theorem holds for open-circuit admittance matrices $Y(p)$.

1.2 A general realizability theorem, applicable to and characterizing completely all finite passive networks, whether having impedance matrices or not, is also proved.

1.3 An effort is made to lay a foundation adequate for the realizability theory of both active and passive multi-terminal devices. To this end, a large part of the paper is devoted to the scrutiny of fundamental properties of networks.

II. INTRODUCTION AND FOREWORD

2.0 Network theory provides direct means for associating with an electrical network a mathematical description which characterizes the behavior of that network. Typically, this results in shifting engineering attention from a detailed, possibly quite intricate, electrical structure to a mathematical entity which succinctly describes the relevant behavior of that structure. An essential feature of this shift in focus is emphasized by the word "relevant": only those terminals of the network which are directly relevant to the problem at hand are considered in the mathematical description. Design work can then be done in terms of constructs relating explicitly to these accessible terminals, the effect of the internal structure being felt only by implication.

The physical origins of these mathematical constructs, and the implications of the internal structure upon them, cannot however be entirely forgotten, for they have mathematical consequences which are not always immediately evident. Until he knows these limitations—imposed upon him by the physical nature or the necessary structural form of the networks he is designing—a design engineer cannot make free use of the mathematical tools that network theory has provided.

We give the name "realizability theory" to that part of network theory which aims at the isolation and understanding of those broad limitations upon network performance, i.e., upon the mathematical constructs which describe that performance—which are imposed by limitations on the network structure. One may also include in the province of realizability theory some of the converse questions: the study of those structural features common to all networks whose performance is limited in some specified way.

Realizability theory would have little content were it not that "per-

formance" here must be construed to mean *performance as viewed from the accessible terminals only*. Were all branch currents and node potentials in a network available to observation, a mathematical statement of performance would be equivalent to stating the full system of differential equations governing these quantities, i.e., equivalent to giving the detailed network diagram.

2.1 With a few important exceptions, the converse kind of problem in realizability theory does not lead to a strict implication from functional limitations to structural features, because the field of equivalent structures for a specified performance is very broad. Typically, it is only by imposing some general *a priori* limitations on structure that further conclusions can be firmly drawn from a functional limitation. In studying this kind of problem one is rapidly led from those basic issues which are clearly part of realizability theory toward general, difficult, and usually unsolved problems of network synthesis. One cannot, and should not, draw a sharp boundary here, but Nature so far has provided a fairly definite one for us, in that most of these problems have proved too difficult of solution.

2.2 The direct realizability problems, the passage from structural properties to functional properties, have been somewhat more tractable. Here, again, there is no clear dividing line between general realizability theory and the sort of design theory in which, for example, one specifies a particular filter structure depending on a limited number of parameters and examines the performance of the structure as a function of these parameters. There is an extensive literature at or near this latter level of generality, most of it relating to filters or filter-like structures (e.g., interstage couplers in amplifiers).

At a more basic level, the limitations on a network's structure which are commonly met in practice are of the following kinds:

a. Limitations on the kind of elements appearing, e.g., to passive networks, networks without coupled coils, networks whose elements have specified parasitics, etc;

b. Limitations on the general form of the network diagram, e.g., to ladder or lattice structures, without limitation to a specified number of elements or parameters.

Here the problems are varied and difficult. We survey briefly the present status of some of them.

2.3 Networks with two accessible terminals, two-poles, are basic in network technology. Fortunately, also, two-poles are unique among networks in that there is always a simple way to describe their perform-

ance. Except for the trivial limiting case of an open circuit, every two-pole has a well-defined impedance, $Z(p)$, a function of the complex frequency parameter p , which describes its performance in a way which is by now well understood. Dually, except for the limiting case of a short circuit, every two-pole has a well-defined admittance function $Y(p)$. Even the limiting cases are tractable: every open circuit has the admittance function $Y(p) \equiv 0$ and every short circuit the impedance function $Z(p) \equiv 0$.

In other words, by exercising his option to speak in terms either of impedance or of admittance, one can always describe the performance of a two-pole by using a single function of frequency.

The descriptive simplicity and practical importance of two-poles led early to a fairly complete realizability theory for them. In 1924 R. M. Foster⁷ gave a function-theoretic characterization of the impedance functions of finite passive two-poles containing only reactances. The corresponding problem for two-poles which are not at all limited as to structure, beyond being finite and passive, was solved by O. Brune² in 1931. The effects of various structural limitations have since been studied by several writers (cf. Darlington,⁶ Bott and Duffin¹³).

2.4 Technology, and the promptings of conscience, have meanwhile urged the study of devices with more than two accessible terminals. Here, however, Nature has been less kind, in that no uniquely simple method is available for describing the performance of such devices as viewed from their terminals.

Indeed, basic network theory has been remiss here, in not even making available a mode of description which is generally applicable—whether simple or not.

W. Cauer⁵ showed that, when one admits ideal transformers among his network components, it is sufficient to study networks which are natural and direct generalizations of two-poles, namely, $2n$ -poles,* for arbitrary values of n . The corresponding natural generalization of the impedance function $Z(p)$ of a two-pole is the impedance *matrix* of a $2n$ -pole: just as one multiplies a scalar current by a scalar impedance to get a scalar voltage, one multiplies a vector current by an impedance matrix to get a vector voltage.

2.41 Not all descriptive difficulties are resolved, however, by considering $2n$ -poles and their impedance or admittance matrices. For the moment, a simple example will suffice to show this: the 2×2 -pole which consists simply of one pair of short-circuited terminals and one pair of

* Defined in Cauer,⁵ and also later here.

open-circuited terminals is a finite passive $2n$ -pole ($n = 2$) which has neither an impedance matrix nor an admittance matrix.

2.42 When one eliminates this kind of descriptive difficulty by fixing his attention only upon $2n$ -poles for which an impedance matrix (or, dually, an admittance matrix) is available, the general realizability problem for finite passive devices is solved. A partial solution, for the case $n = 2$, was published by C. M. Gewertz⁸ in 1933. The solution (Theorem 1.1) of the problem for a general value of n has been announced recently by three authors, independently: Y. Oono,¹⁰ in 1946,* the present author, in 1948,† and M. Bayard,¹ in 1949. The problem for reactive $2n$ -poles is much simpler and was solved by Cauer,³ in 1931.

2.5 Intermediate between the fairly specific problems of filter theory on the one hand and the general realizability theory of multi-terminal devices on the other, lies the study of four-poles as transducers. There is a considerable literature on the realization of transfer functions or transfer impedances under various structural limitations. The basic and extensive work of Bode¹⁴ on active systems belongs also in this category since it is avowedly limited to single-loop structures.

2.6 Beyond the important result that, by sufficiently elaborate conventions, one may avoid the use of transformers in the synthesis of any two-pole, (Bott and Duffin¹³) little in general is known about networks which do not have transformers.

2.7 The present paper is an essay in the realizability theory of devices with many accessible terminals. Ideal transformers are admitted as network elements; indeed, their use is essential. This fact is indicated by the adjective "formal" appearing in the title.

The availability of ideal transformers makes it possible to exploit the simplification noted by Cauer and to consider only networks which are $2n$ -poles in his sense. The aim of the paper, therefore, is to set a foundation for realizability theory which is completely general within this framework.

2.71 The first problem is that of description. We observed above an example—entirely trivial—of a passive four-pole which had neither an impedance nor an admittance matrix. Unfortunately, opportunities

* Date of Japanese publication. This reference, and manuscript of Oono^{10, 11}, were sent by Professor Oono in a personal communication to R. L. Dietzold, who showed them to me in December, 1948, while an early draft of the present paper was in preparation. The recent (1950) American republication of Oono¹⁰ unfortunately omits reference to the original.

† Cf. footnote to 1.1.

for this kind of degeneracy become manifold in multi-terminal devices, and some degree of degeneracy is the rule rather than the exception. Consider an entirely practical example: that of an amplifier chassis from which the tubes have been removed.* Here the degeneracy is essential and intrinsic; it would be highly artificial to regard it as the mere accident of a limiting case. True, given any *particular* degenerate network, there is usually evident a method for representing or describing its behavior. What is needed, however, is a mode of representation which is applicable generally to any $2n$ -pole without *a priori* knowledge of its structure or peculiar degeneracies.

2.72 The mode of representation adopted in this paper, embodied in the notions of general $2n$ -pole (Section 4) and linear correspondence (Section 6), is an obvious one, and so completely general that it solves no problems other than the elemental one for which it was introduced. It provides a definite mathematical construct whose properties one can discuss with mathematical precision. This is all that we ask of it.

Realizability theory begins and ends with the study of these properties. It would be more accurate to say that the notion of general $2n$ -pole describes a particular, but still very large, class of mathematical entities; realizability theory consists in the study of certain subclasses of the whole class of these entities, the particular subclasses being distinguished by special, and to us interesting, properties.

2.73 Despite its avowed aim at generality, the paper is oriented toward the realizability theory of finite passive networks. It ultimately provides a proof of 1.1 and indeed a complete characterization of finite passive $2n$ -poles, however degenerate. This characterization is accomplished in a sequence of postulates, each one delineating a property of general $2n$ -poles, i.e., a subclass consisting of all $2n$ -poles having this property. The class of $2n$ -poles having all of these properties is then identified with the class of $2n$ -poles obtained from finite passive networks.

2.74 If we have succeeded here in our hope to set an adequate foundation for the realizability theory of devices with many terminals, it will be because of the nature and organization of the postulates themselves. They describe what at present seem to be individually significant properties of $2n$ -poles, of progressively greater specificity, which in the aggregate characterize finite passive devices. By eliminating them in various combinations one obtains larger classes of objects. Further re-

* It is exactly this example, and the practical need of an adequate theory for it, which led the author first to study the realizability theory of passive multi-terminal devices.

search alone will tell whether or not one obtains in this way the kinds of device which are significant. For example, one would like general realizability theorems for structures containing vacuum tubes with frequency-independent transconductances, vacuum tubes with non-vanishing transit times, unilateral devices with specified parasitics, etc.

2.75 Actually, the postulates as we have given them are certainly not adequate for such an ambitious program. Exigencies of the presentation have dictated a number of condensations and compromises. It is hoped that the basic ideas are still evident even if not isolated individually in separate and entirely independent postulates. In any event, it is the author's firm belief that the presentation as given is at least illustrative of the kind of approach, and the level of mathematical detail, which will be needed if one is ever to provide a truly adequate realizability theory: a theory which will cover, for example, the broad range of active linear systems which present-day technology allows us to consider.

2.8 Apart from the network theoretic concepts, which must be evaluated by their effectiveness in solving problems—an assessment which is by no means yet complete—this paper is strongly marked by an idiosyncrasy of its author: a consistent and insistent use of geometric ideas and terminology. This is based on the personal experience that linear algebra achieves logical unity and a freedom from encumbering notation when viewed in this way. A general reference covering most of the linear algebra (geometry) required here is P. R. Halmos' elegant monograph⁹.

2.9 For a proof solely of 1.1, which has already been three times proved in the literature,^{1, 10, 11} this paper provides an apparatus which is too cumbersome. There is even a sense in which 1.1 alone provides a characterization of all finite passive devices, for it seems to be generally accepted that, by the use of ideal transformers, any finite passive network can be represented as a network which has an impedance matrix to which is adjoined suitable ideal transformers. Therefore we cannot claim that, in using this cumbrous apparatus to characterize all finite passive $2n$ -poles (including the degenerate ones), we have offered anything not already provided by a simpler proof of 1.1.

Three things may be said in rebuttal. First, we have already emphasized that the apparatus here exhibited was designed for more problems than that to which it is here applied. It is presented in the belief that it will prove of further use.

Second, even in the study of passive networks, it has seemed to the author helpful to look at the manifold things which are *not* passive net-

works. One gets then a clearer view of the unique position occupied by passive devices among all linear systems.

Third, there is a kind of semantic issue here: the assertion that any finite passive *network* (sic) can be put in such a form that 1.1 applies seems to this author to give a kind of circular characterization of such devices. A characterization which did not itself involve the concept of a network seems more satisfying. Logically, there is no circle here, but this is a fact requiring proof. A careful reading of this paper will show that it provides a proof. This particular subtlety does not of itself justify the lengths to which we have gone. It is, however, no longer a subtlety if one wishes to consider devices which do not have a representation in terms of something non-degenerate to which ideal transformers have been added.

2.91 The present Part I of the paper is so organized that at the end of Section 8 the reader is in possession of all of its principal results and its basic ideas. The remaining Sections, 9 through 20, may then be regarded as an Appendix containing the details of proofs. Indeed, Part II will be largely devoted to further details of proof, though there will be there one important idea not mentioned, save casually, in Part I—the idea of degree for a matrix.

In Sections 4 through 11, technical paragraphs have been distinguished from explanatory or heuristic ones by starring the paragraph numeral.

Part II of the paper contains the bulk of the proof of 1.1. This proof is modelled after that of Brune² for the realizability of two-poles. One familiar with the Brune process will probably find Part II readable without extensive reference to Part I.

Let the reader be warned that the Brune process is not a practical one for realizing networks because of its critical dependence upon a difficult minimization and balancing operation. The same criticism applies to the generalized Brune process of Part II.

The Brune process is of theoretical importance because it does realize a network with the minimum number of reactive elements. These facts will be brought to light in Part II.

The proofs of Oono¹⁰ and Bayard¹ are different from ours. That of Oono¹¹ again follows the Brune model.

III. INTRODUCTION TO PART I

3.0 We keep before us first the problem of finding a mathematical description applicable to and characterizing the behavior of all finite pas-

sive networks. Second, we seek to make mathematically precise those ideas which appear to form the basis of general realizability theory. Sections 4 through 7 introduce the immediate mathematical machinery for this. Section 8 states the fundamental realizability theorem and outlines its proof. At this point the reader has had an introduction to the results of the paper. The remainder of the paper is then devoted to the technical details of proof. Beginning with Section 12, the device of starring the technical passages will be dropped.

3.1 Cauer⁵ distinguished precisely the class of networks called $2n$ -poles from the class of all multi-terminal networks. He also showed that, by the use of ideal transformers, any multi-terminal network is equivalent to a network which is a $2n$ -pole (for some n) in his sense. We shall in Section 4 define a class of objects to be called general $2n$ -poles. This class includes all electrical networks which are $2n$ -poles in Cauer's sense. Its definition abstracts the significant properties isolated by Cauer.

For the study, alone, of finite passive networks, this definition is unnecessary, since one can in fact so put the arguments as to deal only with $2n$ -poles which are finite passive networks, and therefore to deal only with concepts already defined in Cauer⁵. The somewhat physical notion of a general $2n$ -pole is a convenient backdrop against which to display the important physical properties of finite passive networks, and, indeed, of networks in general. Having it available, we use it throughout the realizability arguments.

IV. DEFINITION OF GENERAL $2n$ -POLE

4.0* Network theory establishes a correspondence between oriented linear graphs and systems of differential equations. With each node of the graph is associated a potential $E_n = E_n(t)$ and with each oriented branch a current $I_b = I_b(t)$. These potentials and currents are constrained, first by Krichoff's laws, and second by differential equations which depend upon the nature of the branches but not upon the topology of the graph.

4.01* A finite passive network is one whose graph has the following properties:

- (i) There are finitely many nodes, $1, 2, \dots, N$.
- (ii) There are finitely many branches, $1, 2, \dots, B$.

* Technical paragraph as explained in Section 2.91.

- (iii) Let the b -th branch begin at node n_b and end at n'_b . Let $v_b = E_{n_b} - E_{n'_b}$. Then for each b , one of

$$v_b = R_b I_b, \quad R_b > 0 \quad (a)$$

$$I_b = C_b \frac{dv_b}{dt}, \quad C_b > 0 \quad (b)$$

$$v_b = \sum_{b'} L_{bb'} \frac{dI_{b'}}{dt} \quad (c)$$

holds, where the matrix $L_{b,b'}$, is real, symmetric, and semi-definite.

Cauer has shown⁵ how an ideal transformer can be defined as the limiting case of a finite passive network. It is indeed no more nor less ideal than an open circuit ($R_b = \infty$ or $C_b = 0$) or a short circuit ($R_b = 0$ or $C_b = \infty$).

4.02 We seldom deal with networks in the detail which is implicit in (iii) above. We are usually interested in the external characteristics, so to speak, of such networks as viewed from a relatively small number of terminals (nodes). These multi-terminal devices, however, we continue to incorporate into larger network diagrams. It is usually clear how Kirchhoff's laws are to be applied in these cases, and what the differential equations of the resulting system are. We are obliged, however, to make these matters precise before we can deal intelligently with the most general physical properties of networks.

4.1 We have seen the two kinds of constraint that a multi-terminal device imposes on the branch currents and node voltages in a network in which it is incorporated: the topological ones contained in Kirchhoff's laws and the dynamical ones described by differential equations. Correspondingly, there are two aspects to the concept of general $2n$ -pole.

4.11* In its relation to Kirchhoff's laws, a general $2n$ -pole is indicated as an object with n pairs of terminals (T_r, T'_r), $1 \leq r \leq n$. Each terminal can be made a node in an arbitrary finite diagram constructed out of network elements and other general $2m$ -poles, with arbitrary values of m . This diagram is not an oriented linear graph, so we have no basis for the use of Kirchhoff's laws. From it, however, we construct an oriented linear graph, called the *ideal graph* of the diagram, by the following rule:

The nodes of the ideal graph are those of the original diagram. Every

* Technical paragraph as explained in Section 2.91.

oriented branch of the original diagram is repeated in the ideal graph, similarly situated and oriented. Between those nodes which, in the original diagram, were the (T_r, T'_r) of a $2n$ -pole \mathbf{N} , is drawn a branch β_r , called the r -th *ideal branch* of \mathbf{N} , oriented from T_r to T'_r . This is done for each such terminal pair.

Kirchoff's laws now apply to this ideal graph.

4.12* Consider a particular $2n$ -pole \mathbf{N} . Let E_r be the potential of T_r , E'_r that of T'_r . Define

$$v_r(t) = E_r - E'_r.$$

Then $v_r(t)$ is the voltage across the ideal branch β_r so oriented that $v_r(t) \geq 0$ when T_r is positive relative to T'_r . Let $k_r(t)$ represent the current entering T_r . Then $k_r(t) = I_r(t)$, the current in β_r , so $k_r(t)$ is also the current leaving T'_r . This is the force of the notion of ideal branch and the fact which distinguishes a network which is a $2n$ -pole from an arbitrary network with $2n$ terminals.

4.13 For example, the network at (a) of Fig. 1 is not a 2×2 pole because its currents are not constrained to meet the ideal branch requirement. The addition of ideal transformers in either of the ways shown in (b) or (c) of the figure converts it to a 2×2 pole. Of course in a circuit in which the currents are constrained externally—as they would be, for

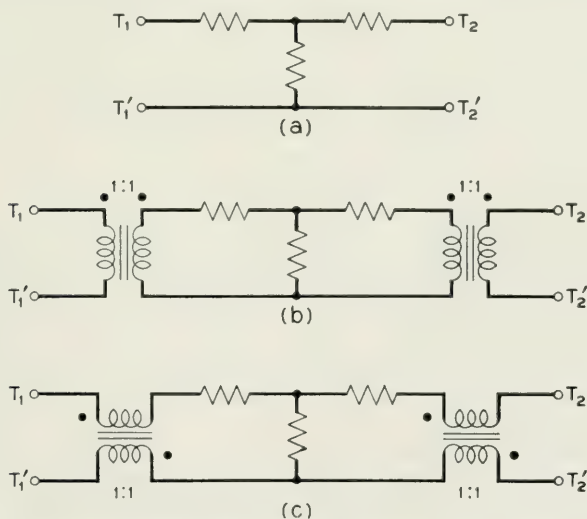


Fig. 1—Conversions of a four pole to a 2×2 pole.

* Technical paragraph as explained in Section 2.91.

example, when the 2×2 pole is driven by separate generators in the two external meshes—these transformers can be eliminated. The definition of $2n$ -pole requires however that in every context the ideal branch concept is valid.

4.2* The second aspect of the concept of general $2n$ -pole is that it imposes some kind of constraint—other than that implied by 4.11 and Kirchhoff's laws—upon the voltages across and currents in its ideal branches. Define the symbols

$$\underline{v} = \underline{v}(t) = [v_1(t), v_2(t), \dots, v_n(t)]$$

and

$$\underline{k} = \underline{k}(t) = [k_1(t), k_2(t), \dots, k_n(t)]$$

as the n -tuples, respectively, of voltages across (T_r , T'_r) and currents into T_r , $1 \leq r \leq n$. These are added and multiplied by scalars by the usual rules of vector algebra. If \underline{v} and \underline{k} represent *simultaneous* values of voltage and current in the $2n$ -pole \mathbf{N} —i.e., values satisfying all the constraints—then we say that \mathbf{N} admits the pair $[\underline{v}, \underline{k}]$.

We say that \mathbf{N} admits \underline{v} if there is a \underline{k} such that \mathbf{N} admits the pair $[\underline{v}, \underline{k}]$. This \underline{k} is said to correspond to \underline{v} . Dually, \mathbf{N} admits \underline{k} if there is a \underline{v} (corresponding to \underline{k}) such that \mathbf{N} admits $[\underline{v}, \underline{k}]$.

The constraints imposed by a general $2n$ -pole \mathbf{N} on voltages and currents are completely described by the totality of pairs $[\underline{v}, \underline{k}]$ which \mathbf{N} admits. We shall *define* a general $2n$ -pole, therefore, as

- (i) a collection of n oriented ideal branches, as in 4.11, and
- (ii) a list of pairs $[\underline{v}, \underline{k}]$ of voltages and currents admitted in these branches.

Hereafter we shall usually drop the adjective "general."

4.21 The definition we have just given is, in a way, a postulational form of an n -dimensional Thevenin's theorem. It postulates that a $2n$ -pole is a thing† which, as far as the outside world is concerned, can be represented by a collection of two-poles β_r , $1 \leq r \leq n$, among which there exists a complicated agreement as to what currents and voltages will be admitted.

4.22 The passive networks (b) and (c) of Fig. 1 define 2×2 poles, because they satisfy 2.01 and clearly permit a complete specification of the admissible pairs $[\underline{v}, \underline{k}]$. Any equivalent network would also specify

* Technical paragraph as explained in Section 2.91.

† In fact, at this level of generality, *any* thing.

the same 2×2 pole, because—by its very equivalence—it would admit the same pairs. The closest association we can make between a $2n$ -pole and a network, then, is to identify the $2n$ -pole with an equivalence class of networks.

4.23 The completely symmetric role played by voltages and currents in this definition of general $2n$ -pole will make it possible to take early advantage of the well-known duality principle of network theory. We shall do so freely.

4.3* We shall call a $2n$ -pole physically realizable if its admissible pairs $[v, k]$ are the solutions of a system of differential equations obtained from a finite passive network, admitting the limiting elements: ideal transformers, open circuits, and short circuits.

V. PHYSICAL PROPERTIES OF NETWORKS

5.0 There are clearly a great many properties of finite passive networks which are not yet possessed by the general $2n$ -poles now introduced. It is instructive to examine these properties physically.

5.1 In the first place, the dynamical constraints (a), (b), and (c) of 4.01 are expressed by linear, time invariant, differential equations. Accordingly, the $2n$ -poles of network theory are:

5.11 Linear, in that the class of admissible pairs $[v, k]$ is a linear space;

5.12 Time invariant, admitting with each $[v(t), k(t)]$ also all $[v(t + \tau), k(t + \tau)]$ for arbitrary τ .

5.2 In the second place, a physical network **N** cannot predict the future, i.e., it cannot respond before it is excited. This can be formalized in terms of the pairs $[v, k]$ admitted by **N**, but to do so would require some digression. The reasons will be seen under 5.7 below.

5.3 We have already mentioned the constraints imposed on voltages and currents in a network by the topology of the network, through the medium of Kirchoff's laws. These constraints have three important properties:

5.31 They are workless, since they are imposed by resistanceless connections, leakless nodes, and, in the formal theory, by ideal transformers.

5.32 Though it seems scarcely necessary to say it, they are the only workless constraints. All other constraints are dynamical and have powers or energies associated with them.

* Technical paragraph as explained in Section 2.91.

5.33 They are frequency independent, that is, holonomic in the sense of dynamics.

5.4 The workless and the dynamical constraints in a physical network are all defined by relations with real coefficients. The space of admissible pairs is then a real linear space.

5.5 The positivities specified in 4.01 are characteristic of passive systems. They correspond to the fact that the power dissipation and the stored energies are all positive.

5.6 By definition, finite passive networks contain finitely many lumped elements. Correspondingly, their resonances and anti-resonances are finite in number.

5.7 We are accustomed to dealing with networks which have, in addition to the properties listed above, a kind of non-degeneracy, in that the list of admissible pairs $[v, k]$ satisfies:

5.71 At least one of v or k can be specified arbitrarily—any real function is admitted;

5.72 When the free number of $[v, k]$ is specified, the other is uniquely determined.

For these non-degenerate networks, the property 5.2 above is easily formalized: if, say, k is determined by v , then

$$v^1(t) = v^2(t) \quad \text{for } t \leq t_0$$

implies

$$k^1(t) = k^2(t) \quad \text{for } t \leq t_0,$$

where $[v^i, k^i]$ are admissible pairs, $i = 1, 2$. The general statement of 5.2 involves this condition and some discussion of the v 's for which **N** admits $[v, 0]$, and the dual notions.

5.8 The reason for speaking in terms of pairs $[v, k]$, instead of in terms of "cause" and "effect," or "impulse" and "response," is hinted at by 5.7 above. For the tacit implications of the cause and effect language completely obscure the fact that 5.71 and 5.72 are properties which are not automatically possessed by electrical networks. In fact, the simple four-pole of 2.41—a pair of unconnected terminals T_1, T'_1 , and a pair of shorted terminals T_2, T'_2 —has neither property, yet it is a perfectly good linear time invariant four pole. Its admissible pairs are

$$[(v_1, 0), (0, k_2)],$$

where v_1 and k_2 are arbitrary real functions of the time

VI. LINEAR CORRESPONDENCES

6.0 In developing the formal properties of $2n$ -poles which are equivalent to the physical ones just listed, it would be instructive to adjoin requirements piecemeal, much in the order given in Section 5. Space does not permit us full enjoyment of this luxury, but the reader will find a rough parallel between Section 5 and the developments of this Section and Section 7.

6.1 It is well known that linear time invariant systems are best studied by the tools of Fourier or Laplace analysis. We make this fact the basis of our first step in characterizing physically realizable $2n$ -poles simply by phrasing our whole discussion in the frequency language. The content of the following paragraph will be obvious enough, but it does define terms to be used later.

6.11* Let v and k , without underscores, represent n -tuples of complex numbers:

$$v = [v_1, v_2, \dots, v_n], \quad (1)$$

$$k = [k_1, k_2, \dots, k_n]. \quad (2)$$

These are to be manipulated by the rules of vector algebra. Let p be a complex number. We shall say that a $2n$ -pole \mathbf{N} admits the pair $[v, k]$ at frequency p , if in the sense of 4.2 \mathbf{N} admits the pair $[\underline{v}, \underline{k}]$ (*with underscores*) where \underline{v} has components

$$v_r(t) = \operatorname{Re}(v_r e^{pt}), \quad 1 \leq r \leq n, \quad (3)$$

and \underline{k} has components

$$k_r(t) = \operatorname{Re}(k_r e^{pt}), \quad 1 \leq r \leq n. \quad (4)$$

Also analogously to 4.2, we say that \mathbf{N} admits v at frequency p if there is a k such that \mathbf{N} admits $[v, k]$ at frequency p , and that this k corresponds to v (at frequency p). Similarly, \mathbf{N} admits k at frequency p if there is a (corresponding) v such that \mathbf{N} admits $[v, k]$ (at p).

6.12* Let \mathbf{V} denote the aggregate of all n -tuples (1), and \mathbf{K} the aggregate of all n -tuples (2). These are then complex linear spaces.

6.2* As our first step toward characterizing realizable $2n$ -poles, let us consider a *linear correspondence* L between \mathbf{V} and \mathbf{K} described by the postulates:

P1. There is a set Γ_L of complex numbers and for each $p \in \Gamma_L$ a list $L(p)$ of pairs $[v, k]$, $v \in \mathbf{V}$, $k \in \mathbf{K}$.

* Technical paragraph as explained in Section 2.91.

P2. If $[v^1, k^1] \epsilon L(p)$ and $[v^2, k^2] \epsilon L(p)$, then

$$[a_1 v^1 + a_2 v^2, a_1 k^1 + a_2 k^2] \epsilon L(p)$$

for any complex numbers a_1, a_2 .

6.21* Given such a linear correspondence L , we can always describe a $2n$ -pole \mathbf{N}_L by:

\mathbf{N}_L admits $[v, k]$ at frequency p if and only if $[v, k] \epsilon L(p)$.

That is, we can always interpret the pairs $[v, k]$ generated by (3) and (4) from the $[v, k] \epsilon L(p)$, for each $p \in \Gamma_L$, as the voltages across and currents in a set of n ideal branches. We call \mathbf{N}_L the $2n$ -pole associated with L .

6.22* We call Γ_L the frequency domain of L (or of \mathbf{N}_L).

6.23 From here on, the words " $2n$ -pole" can with some strain be regarded as suggestive but unnecessary. We in fact deal with linear correspondences—having properties as yet unspecified—and shall show how physical networks can be constructed which admit the pairs $[v, k] \epsilon L(p)$. Actually we use freely the concept of general $2n$ -pole and thereby avoid some elaborate circumlocutions.

6.24* We identify two correspondences L_1 and L_2 as being the same if (i) their frequency domains differ only by a finite set, and (ii) for each p where both are defined the lists $L_1(p)$ and $L_2(p)$ are the same.

6.3 The simplest linear correspondences are those generated by matrices. For example, let $Z(p)$ be an $n \times n$ matrix with, say, elements $Z_{rs}(p)$ which are rational functions of p , $1 \leq r, s \leq n$. Let Γ_L consist of all the values of p at which $Z(p)$ is defined. For $p \in \Gamma_L$, define $L(p)$ as the class of all pairs

$$[v, k] \quad (5)$$

obtained by letting k range over \mathbf{K} , where for each k , v is defined by the matrix equation

$$v = Z(p)k. \quad (6)$$

This kind of matrix equation will be used throughout to symbolize the n component equations

$$v_r = \sum_{s=1}^n Z_{rs}(p)k_s, \quad 1 \leq r \leq n. \quad (7)$$

The list of pairs $L(p)$ defined by (5) clearly satisfies P1 and P2. It can therefore be used to define a $2n$ -pole \mathbf{N}_L . It is easy to see that \mathbf{N}_L

* Technical paragraph as explained in Section 2.91.

in fact is non-degenerate in a sense similar to that of 5.7, for the current amplitudes k can be specified arbitrarily, and the resulting voltage amplitudes v are then fixed by k and p , by (6).

$Z(p)$ is called the impedance matrix of the $2n$ -pole \mathbf{N}_L . It is also sometimes called the open-circuit impedance matrix, because each $Z_{rs}(p)$ is, by (7), the voltage amplitude across (T_r, T'_r) when the current amplitudes at all terminals save (T_s, T'_s) are zero—i.e., when all pairs save the s -th are on open circuit.

6.31 Dually, the pairs

$$[v, Y(p)v]$$

defined by an admittance matrix $Y(p)$ as v ranges over \mathbf{V} define a linear time invariant $2n$ -pole which is non-degenerate.

VII. WORK AND ENERGY

7.0* A linear correspondence satisfying P1 and P2 is something which abstracts the properties of linearity and time invariance. Most of the remaining properties of physical networks involve the mention of work or energy. These concepts enter our picture by way of the scalar product (v, k) between a voltage n -tuple (1) and a current n -tuple (2), of 6.11. This scalar product is defined by

$$(v, k) = \sum_{r=1}^n v_r \bar{k}_r. \quad (1)$$

7.01 If $p = i\omega$, one easily calculates from (3) and (4) of 6.11 that

$$2 \operatorname{Re}(v, k) = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T \left[\sum_{r=1}^n v_r(t) k_r(t) \right] dt.$$

That is, when $p = i\omega$, the real part of $2(v, k)$ measures the average total power dissipated by the system of currents $k_r(t)$ against the driving voltages $v_r(t)$.

When p is not a pure imaginary, the interpretation of the scalar product (v, k) is not so clearly physical as this. The reader will ultimately observe that our significant statements about such products can all be reduced to statements applicable when $p = i\omega$, i.e., when the power interpretation is valid.

7.1* An important concept in what follows is that of the annihilator of a linear manifold (Halmos⁹, par. 16). Let $\mathbf{V}_1 \subseteq \mathbf{V}$ be a linear manifold.

* Technical paragraph as explained in Section 2.91.

Then its annihilator $(\mathbf{V}_1)^0$ is the set of all k such that

$$v \in \mathbf{V}_1 \text{ implies } (v, k) = 0.$$

$(\mathbf{V}_1)^0$ is a linear manifold in \mathbf{K} .

Dually, given $\mathbf{K}_1 \subseteq \mathbf{K}$, $(\mathbf{K}_1)^0$ is the linear manifold of all $v \in \mathbf{V}$ such that

$$k \in \mathbf{K}_1 \text{ implies } (v, k) = 0.$$

The annihilator concept is the analog in our general geometric framework of the idea of orthogonality. It clearly suggests a connection with workless constraints.

7.2* The complex conjugate of an n -tuple v (or k) is defined in the obvious way: if

$$v = [v_1, \dots, v_n]$$

then

$$\bar{v} = [\bar{v}_1, \dots, \bar{v}_n].$$

This conjugation operation clearly has the properties

$$\begin{aligned} \bar{\bar{\xi}} &= \xi \\ \overline{a\xi + b\eta} &= \bar{a}\bar{\xi} + \bar{b}\bar{\eta} \end{aligned} \tag{2}$$

where a and b are scalars and ξ and η are (consistently) elements of \mathbf{V} or \mathbf{K} . Furthermore, at once from (1) of 7.0,

$$\overline{(v, k)} = (\bar{v}, \bar{k}). \tag{3}$$

7.21* A linear manifold will be called real if it contains, with any n -tuple also the conjugate of that n -tuple.

7.22* A real manifold is spanned by real n -tuples. This will be proved in the Appendix, Section 20.

7.23* The annihilator of a real manifold is real. For let \mathbf{K}_1 be real and k^1, \dots, k^r be real n -tuples which span \mathbf{K}_1 . Then if $v \in (\mathbf{K}_1)^0$ every

$$(v, k^s) = 0,$$

and conversely. But then also

$$(\bar{v}, k^s) = \overline{(v, k^s)} = \bar{0} = 0,$$

so $\bar{v} \in (\mathbf{K}_1)^0$.

* Technical paragraph as explained in Section 2.91.

7.3* Given a linear correspondence L , we make several definitions:

$\mathbf{V}_L(p)$ is the set of all $v \in \mathbf{V}$ such that there is a k with $[v, k] \in L(p)$.

$\mathbf{K}_L(p)$ is the set of all $k \in \mathbf{K}$ such that there is a v with $[v, k] \in L(p)$.

$\mathbf{V}_{L0}(p)$ is the set of $v \in \mathbf{V}_L(p)$ such that

$$[v, 0] \in L(p).$$

$\mathbf{K}_{L0}(p)$ is the set of $k \in \mathbf{K}_L(p)$ such that

$$[0, k] \in L(p).$$

7.31* The postulate P2 implies that for each $p \in \Gamma_L$, $\mathbf{V}_L(p)$, $\mathbf{K}_L(p)$, $\mathbf{V}_{L0}(p)$ and $\mathbf{K}_{L0}(p)$ are all linear manifolds.

7.32 $\mathbf{V}_L(p)$, for example, is the set of $v \in \mathbf{V}$ such that \mathbf{N}_L admits v at frequency p .

7.4* We now postulate

P3. There exist fixed linear manifolds $\mathbf{V}_L \subseteq \mathbf{V}$, $\mathbf{K}_L \subseteq \mathbf{K}$ such that

(A) For every $p \in \Gamma_L$, $\mathbf{V}_L(p) = \mathbf{V}_L = (\mathbf{K}_{L0}(p))^0$

(I) For every $p \in \Gamma_L$, $\mathbf{K}_L(p) = \mathbf{K}_L = (\mathbf{V}_{L0}(p))^0$.

7.41* We may henceforth write \mathbf{V}_{L0} , \mathbf{K}_{L0} , for $\mathbf{V}_{L0}(p)$, $\mathbf{K}_{L0}(p)$, knowing that, under P3

$$\mathbf{V}_{L0} = (\mathbf{K}_L)^0,$$

$$\mathbf{K}_{L0} = (\mathbf{V}_L)^0.$$

7.42 Linear correspondences satisfying P3 abstract the properties mentioned in 5.3. The equalities $\mathbf{V}_L(p) = \mathbf{V}_L$, $\mathbf{K}_L(p) = \mathbf{K}_L$ guarantee the frequency-independence of the workless constraints. The equalities $\mathbf{V}_L(p) = (\mathbf{K}_{L0}(p))^0$, $\mathbf{K}_L(p) = (\mathbf{V}_{L0}(p))^0$ in a sense guarantee that the only constraints imposed upon admissible currents and voltages (as opposed to constraints relating currents and voltages) are those which arise from open or short circuits, i.e., are workless.

7.43 An illustrative consequence of P3, for example, is that if L satisfies P3 and if \mathbf{N}_L is such that all of the current amplitudes can be specified arbitrarily, then indeed the voltages are determined by the currents. This will appear as a consequence of 8.1. It is a very general theorem about networks of a kind that this author, at least, has not heretofore encountered.

* Technical paragraph as explained in Section 2.91.

7.5* Continuing toward realizability, we introduce

P4. If $p \in \Gamma_L$, then $\bar{p} \in \Gamma_L$. If $[v, k] \in L(p)$, then $[\bar{v}, \bar{k}] \in L(\bar{p})$.

This postulate embodies most of the reality properties of networks. It has as an immediate consequence the

7.51* *Lemma*: If L satisfies P1, P2, P3, and P4, then all of

$$\mathbf{V}_L, \mathbf{V}_{L^0}, \mathbf{K}_L, \mathbf{K}_{L^0}$$

are real.

Proof: By P4, $v \in \mathbf{V}_L(p) = \mathbf{V}_L$ implies $\bar{v} \in \mathbf{V}_L(\bar{p}) = \mathbf{V}_L$. Hence \mathbf{V}_L is real. Then $\mathbf{K}_{L^0} = (\mathbf{V}_L)^0$ is real, and dually.

7.6* The three remaining postulates on L refer to scalar products. They are concerned with the energy questions related to passivity, rather than with the workless constraint questions.

P5. If $[u, j] \in L(p)$ and $[v, k] \in L(p)$, and if

(A) u and v are real, or if

(I) j and k are real,

then

$$(u, k) = (v, j).$$

7.61 This is the property which provides the reciprocity law. In its presence, the relations in P3 may be weakened to

$$\mathbf{V}_L(p) = \mathbf{V}_L \supseteq (\mathbf{K}_{L^0}(p))^0,$$

$$\mathbf{K}_L(p) = \mathbf{K}_L \supseteq (\mathbf{V}_{L^0}(p))^0.$$

This fact will appear as a consequence of the lemma of Section 12.

7.7* *Lemma*: A consequence of P2 and P3(A) is that if

$$[v, k_r] \in L(p), \quad r = 1, 2,$$

then for any $u \in \mathbf{V}_L$,

$$(u, k_1) = (u, k_2).$$

For by P2 we have that

$$[v - v, k_1 - k_2] = [0, k_1 - k_2] \in L(p),$$

hence $k_1 - k_2 \in \mathbf{K}_{L^0}$. Then however, by P3(A), $u \in \mathbf{V}_L$ implies $u \in (\mathbf{K}_{L^0})^0$, so

* Technical paragraph as explained in Section 2.91.

that

$$0 = (u, k_1 - k_2) = (u, k_1) - (u, k_2).$$

Q.E.D. A dual result follows from P3(I).

7.71* The result of 7.7 above means that the scalar product (v, k) is fixed by v alone when we know that $[v, k] \in L(p)$. This means that, given $v \in \mathbf{V}_L$, there is a unique function $F_v(p)$ defined for $p \in \Gamma_L$ by

$$F_v(p) = \overline{(v, k)}$$

where $[v, k] \in L(p)$. Dually,

$$J_k(p) = (v, k)$$

is defined for each fixed $k \in \mathbf{K}_L$.

7.72* (P6.) The complement of Γ_L is finite and

(I) For each $v \in \mathbf{V}_L$, $F_v(p)$ is rational

(A) For each $k \in \mathbf{K}_L$, $J_k(p)$ is rational.

7.73* (P7.) (A) $\operatorname{Re}(p) \geq 0$ implies $\operatorname{Re}(F_v(p)) \geq 0$

(I) $\operatorname{Re}(p) \geq 0$ implies $\operatorname{Re}(J_k(p)) \geq 0$.

VIII. THE FUNDAMENTAL REALIZABILITY THEOREM

8.0* We can now state our fundamental realizability theorem: If a linear correspondence L satisfies P1, \dots , P7, the associated $2n$ -pole \mathbf{N}_L is physically realizable. Conversely, given a physically realizable $2n$ -pole \mathbf{N} , the associated linear correspondence satisfies P1, \dots , P7.

8.01 Actually, the postulates P1, \dots , P7 are not unique nor even entirely independent. Many changes may be rung on them. We indicated one above. At the expense of apparent asymmetry, the (A) or (I) portions, in various combinations, can be deleted or weakened. We shall not pursue this subject further at this point, but must come back to it in Section 12.

8.02 We close this Section by outlining the proof of 8.0. The details are then contained in the remainder of the paper.

8.03 The proof that P1 through P7 are necessary for physical realizability will be a direct one: it will be shown that, considered individually, each network branch and each ideal transformer satisfies the postulates.

* Technical paragraph as explained in Section 2.91.

By an application of Kron's method (described by Synge¹²), it will then be shown that the imposition of Kirchoff's laws preserves the postulates. This work is most efficiently performed after the full machinery of the sufficiency proofs is available, and will be done in Section 19.

8.04 The sufficiency of P1 through P7 can be deduced—and we will do so—from the lemmas to be quoted below. Apart from Section 19 on necessity, the remainder of the paper is devoted to the proofs of these lemmas.

8.1* *Lemma:* If L is a linear correspondence satisfying P1, P2, P3, and P4, then there exists a fixed real nonsingular matrix W such that

8.11 The list $L_w(p)$ of all pairs†

$$[W^{-1}v, W'k],$$

where $[v, k] \in L(p)$, describes a linear correspondence L_w satisfying P1, P2, P3, and P4.

8.12 The $2n$ -pole $\mathbf{N}_w (= \mathbf{N}_{L_w})$ associated with L_w consists of

- (i) Some number r of open-circuited terminal pairs $(T_1, T'_1), \dots, (T_r, T'_r)$,
- (ii) Some number s of short-circuited terminal pairs $(T_{n-s+1}, T'_{n-s+1}), \dots, (T_n, T'_n)$,
- (iii) A set of $m = n - r - s$ terminal pairs $(T_{r+1}, T'_{r+1}), \dots, (T_{r+m}, T'_{r+m})$.

8.13 Either $m = 0$, or the terminal pairs in (iii) are those of a $2m$ -pole \mathbf{N}_1 which has a nonsingular impedance matrix $Z_1(p)$.

This lemma, and the following, will be proved in 13.2.

8.2* *Lemma:* If L satisfies P5, P6, and P7, then $Z_1(p)$ is a positive real‡ matrix, that is, $Z_1(p)$ satisfies (i), \dots , (iv) of 1.1.

8.3* *Lemma:* If a $2m$ -pole \mathbf{N}_1 has a positive real impedance matrix, then \mathbf{N}_1 is physically realizable.

This is the sufficiency half of the matrix realizability theorem 1.1. Part II will be devoted to its proof.

8.4* *Lemma:* If \mathbf{N}_w is physically realizable, then \mathbf{N} can be constructed from it by the use of ideal transformers.

This is Cauer's Transformation Theorem⁵ about which we shall say more in Section 9.

* Technical paragraph as explained in Section 2.91.

† W^{-1} and W' are respectively the reciprocal and the transpose of W .

‡ Gewertz's terminology³, by now traditional.

8.5* The sufficiency half of 8.0 is now clear. By 8.2 and 8.3, \mathbf{N}_1 is physically realizable. Clearly then \mathbf{N}_w is, simply by the adjunction of the necessary open and short circuits. Finally \mathbf{N} is by Cauer's theorem, 8.4.

8.6* We can see now how to prove the necessity of positive reality for the realizability of a positive real matrix $Z(p)$. This is the necessity half of the matrix theorem 1.1. Let $Z(p)$ be the matrix of a realizable \mathbf{N} . Then \mathbf{N} has an associated linear correspondence L satisfying P1, \dots , P7, by the necessity half of 8.0. The pairs of L are the pairs

$$[Z(p)k, k]$$

generated as k ranges over all n -tuples. By definition, then, the pairs of L_w are

$$[W^{-1}Z(p)k, W'k].$$

As k ranges over all n -tuples, the nonsingularity of W implies that $W'k$ does also. Let $U = W^{-1}$. Then the pairs above are the same as

$$[UZ(p)U'k, k]$$

as k ranges over all n -tuples. Hence L_w has the impedance matrix $UZ(p)U'$, where $U = W^{-1}$ is real and nonsingular. Because L_w has an impedance matrix, $r = 0$ in 8.12.

Now by 8.1 and 8.2, $Z_1(p)$ is positive real and the matrix $UZ(p)U'$ of L_w is just $Z_1(p)$ bordered by s rows and columns of zeros. It is then easy to see that $UZ(p)U'$ is positive real, and finally also that $Z(p)$ is. These last two facts will be proved formally in Section 16.

IX. CAUER'S TRANSFORMATION THEOREM

9.0 Cauer's transformation theorem⁵ is the cornerstone of formal realizability theory. In one form, the theorem reads:

9.1* Let $Z(p)$ be the impedance matrix of a physically realizable $2n$ -pole \mathbf{N} . Let U be a real, constant, nonsingular matrix. Then

$$UZ(p)U' \tag{1}$$

is again the impedance matrix of a physically realizable $2n$ -pole, \mathbf{N}_U . \mathbf{N}_U can be constructed from \mathbf{N} by the use of ideal transformers.

9.2* A superficial generalization of this theorem can be obtained at once from Cauer's proof. It asserts that if \mathbf{N} is physically realizable and is described by the linear correspondence L , then there is a physically realizable $2n$ -pole \mathbf{N}_w , obtainable from \mathbf{N} by the use of ideal trans-

* Technical paragraph as explained in Section 2.91.

formers, which is described by the linear correspondence L_w whose pairs at each p are the pairs

$$[W^{-1}v, W'k], \quad (2)$$

where $[v, k] \in L(p)$.

We refer to Cauer⁵ for the proof. It is straightforward.

9.21 We shall use the second form (9.2) of Cauer's theorem in our realization process. Notice that it is in a sense a "physical" theorem, about the way one physical network is related to another. It is used in this way: we shall always solve a realizability problem by finding some network \mathbf{N} which is easily realized, and then a W such that \mathbf{N}_w , which is now realizable, provides a solution to the given problem.

9.22* We shall call the $2n$ -pole \mathbf{N}_w a Cauer equivalent of \mathbf{N} .

9.3 Although Cauer's theorem will be applied, in a sense, only *a posteriori*, its effect is fundamental. For it implies that formal physical realizability is a property of matrices which is invariant under the operation (1) or a property of correspondences which is invariant under (2). There is an extensive classical literature on the properties of matrices invariant under operations like that of (1), and the effect of Cauer's theorem is to make these results all available to formal realizability theory.

9.31* It is worth observing here that we are already well set up to use Cauer's theorem:

Lemma: If L is a linear correspondence satisfying P1, \dots , P7, then the correspondence L_w of 9.2 also satisfies P1, \dots , P7.

Proof: Let $M = L_w$. P1 and P2 for M are obvious, with $\Gamma_M = \Gamma_L$. By definition of M ,

$$\begin{aligned} \mathbf{V}_M(p) &= W^{-1}\mathbf{V}_L(p) = W^{-1}\mathbf{V}_L \\ \mathbf{K}_M(p) &= W'\mathbf{K}_L(p) = W'\mathbf{K}_L \\ \mathbf{V}_{M0}(p) &= W^{-1}\mathbf{V}_{L0}(p) = W^{-1}\mathbf{V}_{L0} \\ \mathbf{K}_{M0}(p) &= W'\mathbf{K}_{L0}(p) = W'\mathbf{K}_{L0} \end{aligned}$$

where $W^{-1}\mathbf{S}$ for a manifold \mathbf{S} consists of all n -tuples $W^{-1}v$, where $v \in \mathbf{S}$. Hence in P3,

$$\begin{aligned} \mathbf{V}_M(p) &= \mathbf{V}_M = W^{-1}\mathbf{V}_L \\ \mathbf{K}_M(p) &= \mathbf{K}_M = W'\mathbf{K}_L \end{aligned}$$

for fixed manifolds \mathbf{V}_M , \mathbf{K}_M as defined.

* Technical paragraph as explained in Section 2.91.

Now if $v \in \mathbf{V}_{L0}$, then

$$(v, k) = 0$$

for every $k \in \mathbf{K}_L = (\mathbf{V}_{L0})^0$. Then, however, by direct calculation from Section 7.0,

$$(W^{-1}v, W^*k) = 0,$$

where W^* is the adjoint, i.e. transposed conjugate matrix of W . But because W is real, $W^* = W'$. Hence if $v \in \mathbf{V}_{L0}$, then

$$(W^{-1}v, k) = 0$$

for every $k \in \mathbf{K}_L = \mathbf{K}_M$. Hence

$$\mathbf{K}_M = (W^{-1}\mathbf{V}_{L0})^0 = (\mathbf{V}_{M0}(p))^0.$$

By this and its dual, P3 is completed for M .

The remaining postulates for M follow from those for L by the simple equality

$$(v, k) = (W^{-1}v, W'k)$$

already established, combined with $\Gamma_M = \Gamma_L$.

9.32 For fixed $Z(p)$, the matrices (1), as U ranges over a group, form an equivalence class. Classical matrix theory treats of such equivalence classes. This author's predilection is to regard this theory from a geometrical point of view. In part this prejudice may be justified by the ease with which that slightly more general object, a linear correspondence, can be treated by geometrical methods. In any event we shall begin our program of proofs with a brief introduction to the geometrical approach.

X. GEOMETRICAL PRELIMINARIES

10.0* We now wish to consider \mathbf{V} and \mathbf{K} as complex n -dimensional linear spaces[†] respectively of voltage vectors v and current vectors k . The distinction here is in point of view. A vector v is regarded as an absolute geometrical object; an n -tuple $[v] = [a_1, \dots, a_n]$ is regarded as a set of coordinates for the vector v , relative to some coordinate basis. Given a fixed coordinate basis, there is a one-to-one correspondence between vectors v and the n -tuples $[v]$ which represent them in that basis, a correspondence which preserves the operations of vector algebra.

* Technical paragraph as explained in Section 2.91.

† For a reference concerning the ideas in this section, see Halmos⁹, Chapters I and II.

10.01 The effect of attaching a geometric identity to vectors, rather than to n -tuples, is to make it possible to choose coordinate bases freely and as convenient, without elaborate constructions or even interpretations. We can then discuss properties of n -tuples (and other objects, e.g. matrices) which are invariant under the kind of operations exemplified by (1) and (2) of Section 9 as *properties of a single geometric object*, rather than as properties shared by an extensive class of concrete objects which are converted into each other by the group of operations.

10.1 This change in point of view need not change formally anything we have said to date; it simply erects a conceptual superstructure, or provides a conceptual foundation, depending on the reader's personal attitude.

We shall support this statement by going through the important ideas of Sections 4, 6, and 7 and examining their geometrical meanings or counterparts. It is convenient to consider first and at some length the notions of scalar product and complex conjugate. The geometric structure will then be complete enough to permit a rapid survey of the remaining ideas.

10.11* The geometrical counterpart of the scalar product introduced in 7.0 is a numerically valued function $\sigma = \sigma(v, k)$ of two vector variables. Its first argument v ranges over \mathbf{V} and its second argument k ranges over \mathbf{K} . The function $\sigma(v, k)$ is linear in v and conjugate linear in k :

$$\begin{aligned}\sigma(au + bv, k) &= a\sigma(u, k) + b\sigma(v, k), \\ \sigma(v, ak + b\ell) &= \bar{a}\sigma(v, k) + \bar{b}\sigma(v, \ell).\end{aligned}\tag{1}$$

We denote this function $\sigma(v, k)$ by the simple bracket notation (v, k) .

10.12 With this scalar product, the geometry of \mathbf{V} and \mathbf{K} is that of a space \mathbf{K} and the space $\mathbf{K}^* = \mathbf{V}$ of conjugate linear functionals over \mathbf{K} . This is analogous to the real geometry of space and conjugate space discussed at length in Halmos⁹. In fact, in the introduction to Chapter III of Halmos⁹, the modifications introduced by the conjugate linearity of (v, k) over \mathbf{K} are treated in detail.

10.13* Because of its importance, we quote here a paraphrase of the results covered in Halmos⁹, par. 12.

(i) If $f(v)$ is any numerically valued homogeneous linear function of $v \in \mathbf{V}$, then there is a unique vector $k_f \in \mathbf{K}$ such that

$$f(v) = (v, k_f)$$

for all $v \in \mathbf{V}$.

* Technical paragraph as explained in Section 2.91.

(ii) If $g(k)$ is any numerically values homogeneous conjugate-linear function of $k \in \mathbf{K}$ (i.e., if $\overline{g(k)}$ is linear in k) then there is a unique $v_\theta \in \mathbf{V}$ such that

$$g(k) = (v_\theta, k)$$

for all $k \in \mathbf{K}$.

10.2* The annihilator $(\mathbf{V}_1)^0$ of a manifold $\mathbf{V}_1 \subseteq \mathbf{V}$ is, as in 7.1, the set of all $k \in \mathbf{K}$ such that

$$v \in \mathbf{V}_1 \text{ implies } (v, k) = 0.$$

10.21* It is shown in Halmos⁹ that to each basis v^1, \dots, v^n in \mathbf{V} there exists a unique dual basis k^1, \dots, k^n in \mathbf{K} such that

$$(v^r, k^s) = \delta_{rs}, \quad (2)$$

where δ_{rs} is the Kronecker symbol: $\delta_{rs} = 0$ if $r \neq s$, $\delta_{rr} = 1$, $1 \leq r, s \leq n$.

10.22 If

$$\begin{aligned} [v] &= [a_1, \dots, a_n] \\ [k] &= [b_1, \dots, b_n] \end{aligned} \quad (3)$$

are the n -tuples representing v and k relative to a pair of dual bases, then it is easily computed from (1) and (2) that

$$(v, k) = \sum_{r=1}^n a_r \bar{b}_r. \quad (4)$$

Therefore the concrete scalar product of 7.0 is indeed the geometric scalar product here considered, when we restrict our pairs of bases in \mathbf{V} and \mathbf{K} always to be dual in the sense of (2).

10.23* We shall use the words “coordinate frame” or simply “frame” to denote a pair of dual bases in \mathbf{V} and \mathbf{K} . Any basis in \mathbf{V} (or \mathbf{K}) specifies a frame by the uniqueness result quoted above.

10.24 We shall henceforth deal always with coordinate frames, in fact, ultimately, real coordinate frames, rather than arbitrary pairs of bases. This means in classical language that we are considering as “geometrical properties” all properties which are preserved under the group of linear transformations which leave the bilinear form (4) invariant. The properties related to physical realizability will turn out to be invariant only under the subgroup of real linear transformations preserving (4).

* Technical paragraph as explained in Section 2.91.

10.3* Conjugation is an operation which to each $v \in \mathbf{V}$ associates a vector \bar{v} uniquely determined by v with the properties

$$\begin{aligned}\bar{\bar{v}} &= v, \\ \overline{(au + bv)} &= \bar{a}\bar{u} + \bar{b}\bar{v},\end{aligned}\tag{5}$$

where a and b are any complex numbers and \bar{a} , \bar{b} their conjugates.

10.31* Given any such conjugation operation in \mathbf{V} , and given any $k \in \mathbf{K}$, define a function $g_k(v)$ by

$$g_k(v) = (\bar{v}, \bar{k})\tag{6}$$

for $v \in \mathbf{V}$. Then $g_k(v)$ is linear in v , by (5) above and (1) of 10.11. Therefore, by 10.13, there is a unique vector $\bar{k} \in \mathbf{K}$ such that

$$g_k(v) = (v, \bar{k}).\tag{7}$$

10.32* Directly from (1) of 10.11 and (6) above, if $j = ak + b\ell$, then

$$g_j(v) = ag_k(v) + bg_\ell(v).$$

From (7), therefore

$$(v, \bar{j}) = a(v, \bar{k}) + b(v, \bar{\ell})$$

for all $v \in \mathbf{V}$. Comparing this with (1) of 10.11, we see that

$$\bar{j} = \bar{a}\bar{k} + \bar{b}\bar{\ell}.\tag{8}$$

The second item of (5) above then holds for vectors $k \in \mathbf{K}$.

That $\bar{\bar{k}} = k$ follows easily: We have from (6) and (7), written for the vector \bar{k} , that

$$(\bar{\bar{v}}, \bar{\bar{k}}) = (v, \bar{k}).\tag{9}$$

We also have, by writing (6) and (7) for vectors \bar{v} and k that

$$(\bar{\bar{v}}, k) = (\bar{v}, \bar{k}).$$

Taking complex conjugates of these two numbers, and using $\bar{\bar{v}} = v$ from (5), we have

$$(v, k) = \overline{(\bar{v}, \bar{k})}.\tag{10}$$

Then (9) and (10), which hold for all $v \in \mathbf{V}$, identify k and $\bar{\bar{k}}$ by 10.13.

10.34* We have now showed in (5), (8) and (10) that this complex conjugate satisfies the formal properties of the conjugate for n -tuples introduced in 7.2.

* Technical paragraph as explained in Section 2.91.

10.35. The abstract scalar product of 10.11 turned out in the end to be no more than the concrete one of 7.0 when we restrict our attention to n -tuples derived from vectors by the use of coordinate frames. In a similar way, it is not hard to show that there always exists a coordinate frame in which the abstract conjugation now introduced has the form of 7.2. This will be done in the Appendix (20.2).

10.36* Our need for writing out the components of vectors has now almost vanished. Henceforth we shall use subscripts to denote particular vectors, e.g. v_1 , rather than components.

10.4* A vector will be called real if it is equal to its own conjugate. A manifold will be called real if it contains with each vector also the conjugate of that vector. \mathbf{V} and \mathbf{K} are then real. A basis will be called real if it is made up of real vectors, and a frame will be called real if its bases are real. Any frame in terms of which our conjugation operation takes the form of 7.2 is real by definition because its basis vectors *in that frame* have components which are 0 or 1. The vector 0 is real, similarly.

10.41* The basis dual to a real basis is real, for if

$$(v_r, k_s) = \delta_{rs},$$

then by (10) of 10.3 and the hypothesis that $v_r = \bar{v}_r$, we have

$$(v_r, \bar{k}_s) = \delta_{rs} = \delta_{rs}$$

so the \bar{k}_s satisfy the same equations as the k . The uniqueness of the basis dual to v_1, \dots, v_r then proves that $\bar{k}_s = k_s, 1 \leq s \leq n$.

10.42* Any vector v can be written

$$v = v_1 + iv_2$$

where v_1 and v_2 are real. Namely

$$v_1 = \frac{1}{2} (v + \bar{v}),$$

$$v_2 = \frac{1}{2i} (v - \bar{v}).$$

10.5* It is shown in Halmos⁹, par. 34, that if $v \in \mathbf{V}$, $k \in \mathbf{K}$ are represented by $[v]$, $[k]$ in some coordinate frame, and by $[v]_1$, $[k]_1$ in some other frame, then there is a nonsingular matrix $[W]$, which (a) depends only upon the

* Technical paragraph as explained in Section 2.91.

two frames, and (b) relates these n -tuples as follows:

$$\begin{aligned} [v]_1 &= [W]^{-1}[v], \\ [k]_1 &= [W]^*[k]. \end{aligned} \quad (11)$$

It is easy to show that if $[W]$ has real elements, so that $[W]^* = [W]'$, then the two frames involved above are either both real, or else neither is real. Also, conversely, if both frames are real, then necessarily the $[W]$ of (11) has real elements and $[W]^* = [W]'$.

10.6* Some further important geometrical notions must be mentioned before we proceed.

If \mathbf{V}_1 and \mathbf{V}_2 are disjoint linear manifolds in \mathbf{V} —i.e. linear manifolds having in common only the single vector $\mathbf{0}$ —we write

$$\mathbf{V}_1 \oplus \mathbf{V}_2$$

for the linear manifold consisting of all vectors $v = v_1 + v_2$, where $v_i \in \mathbf{V}_i$, $i = 1, 2$. The circle around the plus sign is used to denote the disjointness of \mathbf{V}_1 and \mathbf{V}_2 .

It is shown in Halmos⁹, par. 19, that if

$$\mathbf{V} = \mathbf{V}_1 \oplus \mathbf{V}_2 \quad (12)$$

then

$$\mathbf{K} = \mathbf{K}_1 \oplus \mathbf{K}_2, \quad (13)$$

where $\mathbf{K}_1 = (\mathbf{V}_2)^0$, $\mathbf{K}_2 = (\mathbf{V}_1)^0$ and the dimension of \mathbf{K}_i is equal to that of \mathbf{V}_i , $i = 1, 2$. We call (13) the decomposition dual to (12). We sometimes write $\mathbf{K}_i = \mathbf{V}_i^*$ to denote the \mathbf{K}_i dual to \mathbf{V}_i in the decomposition (13). It is shown in Halmos⁹, loc. cit., that there exists a basis v_1, \dots, v_n in \mathbf{V} and its dual k_1, \dots, k_n in \mathbf{K} such that, if r is the dimension of \mathbf{V}_1 ,

$$\begin{aligned} v_1, \dots, v_r &\text{ is a basis for } \mathbf{V}_1 \\ v_{r+1}, \dots, v_n &\text{ is a basis for } \mathbf{V}_2 \\ k_1, \dots, k_r &\text{ is a basis for } \mathbf{K}_1 \\ k_{r+1}, \dots, k_n &\text{ is a basis for } \mathbf{K}_2. \end{aligned} \quad (14)$$

Furthermore, if v_1, \dots, v_n is any basis in \mathbf{V} satisfying the first half of (14), its dual basis satisfies the second half, and dually.

We shall show in the Appendix that if any one of \mathbf{V}_1 , \mathbf{V}_2 , \mathbf{K}_1 , or

* Technical paragraph as explained in Section 2.91.

\mathbf{K}_2 is real, then they all are, and that in this case the bases (14) can be chosen to be real.

Similar considerations apply to decompositions into more summands: if

$$\mathbf{V} = \mathbf{V}_1 \oplus \mathbf{V}_2 \oplus \cdots \oplus \mathbf{V}_m$$

then

$$\mathbf{K} = \mathbf{K}_1 \oplus \mathbf{K}_2 \oplus \cdots \oplus \mathbf{K}_m,$$

where

$$\mathbf{V}_i^* = \mathbf{K}_i = \bigcap_{j \neq i} \mathbf{V}_j^0 = \left(\sum_{j \neq i} \mathbf{V}_j \right)^0.$$

XI. GEOMETRICAL CORRESPONDENCES

11.0 With the geometry of \mathbf{V} and \mathbf{K} now in hand, we consider the geometric aspects of our network theoretic concepts.

The definition in Section 4 of general $2n$ -pole describes a concrete thing and stands unaltered in our geometric view. The definitions in 6.11 of the terminology typified by “ \mathbf{N} admits $[v, k]$ at frequency p ” are unchanged except that we should now explicitly indicate that we are discussing concrete n -tuples of complex numbers by placing brackets around the vector symbols, thus: $[v]$, $[k]$. In other words, a $2n$ -pole is described by a concrete relation between n -tuples.

11.1* All of the postulates P1, \cdots , P7 are stated in a language which now has been given an absolute geometric meaning. In this meaning, P1 and P2 describe a *geometrical linear correspondence* between vectors $v \in \mathbf{V}$ and $k \in \mathbf{K}$. This is the geometric counterpart of the concrete notion of a linear correspondence between n -tuples.

11.11 An impedance matrix, as in 6.3, describes a particularly tightly knit linear correspondence, namely a linear function from \mathbf{K} to \mathbf{V} . The geometrical counterpart is an *impedance operator* which for each p is by definition a linear homogeneous function which assigns to each vector $k \in \mathbf{K}$ a unique $v = Z(p)k \in \mathbf{V}$. That is: an operator is a functional relationship between vectors and as such has a geometric identity.

11.12 It is easy to prove† that, given an impedance operator $Z(p)$, and given any coordinate bases in \mathbf{V} and \mathbf{K} respectively, there is a matrix $[Z(p)]$, with elements $Z_{rs}(p)$, $1 \leq r, s \leq n$, such that relative to these bases the coordinates k_s of a vector k and the coordinates v_r of $v = Z(p)k$ are related by (7) of 6.3. We call $[Z(p)]$ the matrix of $Z(p)$

* Technical paragraph as explained in Section 2.91.

† Cf. Halmos⁹, par. 26.

relative to the given pair of bases. A strong analog of this observation is contained in the following lemma.

11.13* *Lemma:* (i) Let L be a geometrical linear correspondence. Fix any real coordinate frame and let $[L]$ be the linear correspondence whose paired n -tuples are

$$[[v], [k]],$$

where

$$[v], [k] \in L(p).$$

(ii) Alternatively, let $[L]$ be a (concrete) linear correspondence between n -tuples. Interpret the n -tuples related by $[L]$ as representing vectors in some real coordinate frame. Let L be the geometrical correspondence whose pairs, expressed as n -tuples in this frame, are those of the concrete correspondence $[L]$.

In either case, (i) or (ii), the geometric correspondence L satisfies the geometric postulates P1, \dots , P7 if and only if the concrete correspondence $[L]$ satisfies the concrete forms of these postulates.

The proof of this lemma consists essentially in reading the postulates carefully. We shall not reproduce it.

11.2 Our position is now this: We have on the one hand geometrical objects, vectors v, k , operators $Z(p), Y(p)$, and geometrical correspondences L . On the other hand, we have concrete n -tuples $[v], [k]$, matrices $[Z(p)], [Y(p)]$, and linear correspondences $[L]$. Given any pair of bases in \mathbf{V} and \mathbf{K} , in particular, given any coordinate frame, each geometric object generates a corresponding concrete object which represents it relative to those bases or that frame. Conversely, given a concrete object $[\xi]$, we can choose a frame in \mathbf{V} and \mathbf{K} and find that geometric object ξ whose coordinates in the chosen frame are given by $[\xi]$.

11.21* The concrete object, linear correspondence, defines a linear time-invariant $2n$ -pole by 6.21. To complete the picture, we might say that a geometrical correspondence L defines a *Cauer class* of $2n$ -poles.

11.22* This terminology is motivated by the following observation: if $[L]$ and $[L]_1$ are linear correspondences representing L in two distinct real frames, then there exists a real nonsingular matrix $[W]$ relating the

$$[[v], [k]] \in [L](p)$$

and the

$$[[v]_1, [k]_1] \in [L]_1(p)$$

* Technical paragraph as explained in Section 2.91.

by the formulas of 10.5. This means that $[L]$ and $[L]_1$ are related like the $[L]$ and $[L_w]$ of 9.2. The $2n$ -pole associated with $[L]_1$ therefore is a Caue equivalent of that associated with $[L]$.

11.23 The observation of 11.22, combined with (ii) of 11.13, gives an alternative proof of 9.31. This proof is deceptively free of calculation, but of course the calculations are concealed in the extensive geometrical developments of Section 10, many of which are there offered on faith.

XII. THE FUNDAMENTAL LEMMA

12.0 This section is devoted to the statement, and the proof in part, of a lemma which, on the face of it, looks like an exercise in manipulating the postulates. In fact, the content of the lemma, and most of the details of its proof, are essential in what follows. To postpone them would force us into needless duplication of effort.

Lemma: Let L be a geometrical linear correspondence satisfying P1, P2, P4, P5(I), P6(I), P7(I) and the following weak form of P3(I):

P3'(I): If $p \in \Gamma_L$, then $\mathbf{K}_L(p) = \mathbf{K}_L \supseteq (\mathbf{V}_{L0}(p))^0$.

Then there is a frequency domain $\Gamma'_L \subseteq \Gamma_L$, differing from Γ_L by a finite set, such that L satisfies all of the postulates for $p \in \Gamma'_L$.

The statement of the dual result is evident and will be omitted.

The proof that L satisfies P3 will be given in this section. Verification of the remaining postulates will follow in paragraph 16.6.

We assume that the properties of positive real (PR) functions are known. They are summarized for later use in Section 15. We make occasional advance references thereto.

To the proof:

12.01 First, \mathbf{K}_L is a real manifold and for $p \in \Gamma_L$

$$\mathbf{K}_L \subseteq (\mathbf{V}_{L0}(p))^0. \quad (1)$$

This, with P3'(I), gives P3(I) for L .

Proof: \mathbf{K}_L is real, as in 7.51. Consider now a $p \in \Gamma_L$ and a $v \in \mathbf{V}_{L0}(p)$; then $[v, 0] \in L(p)$. Consider any real $j \in \mathbf{K}_L$; then there is a $u \in \mathbf{V}_L(p)$ such that $[u, j] \in L(p)$. Now 0 and j are real. Hence by P5(I)

$$(v, j) = (u, 0) = 0.$$

Therefore any real $j \in \mathbf{K}_L$ has a vanishing scalar product with every $v \in \mathbf{V}_{L0}(p)$. Since \mathbf{K}_L is real, it is spanned by real j and (1) follows.

12.1 By the dual of 7.7, if we know that

$$[v, k] \in L(p),$$

then the value of (v, k) is determined by k . This makes it possible to state P6(I) and P7(I) for L (we take P6(I) to include the hypothesis that Γ_L has a finite complement).

12.11 If $k \in \mathbf{K}_L$, then $J_k(p)$ is PR.

Proof: if k is real then

$$\overline{J_k(p)} = \overline{(v, k)} = (\bar{v}, k), \quad (2)$$

where, of course, $[v, k] \in L(p)$. Then however $[\bar{v}, k] \in L(\bar{p})$, by P4. Hence by 12.1, (2) gives us

$$\overline{J_k(p)} = J_k(\bar{p}).$$

From this and P6(I), P7(I) we conclude that $J_k(p)$ is PR for any real $k \in \mathbf{K}_L$.

Now, given any $k \in \mathbf{K}_L$, we have $\bar{k} \in \mathbf{K}_L$ by 12.01. Then

$$k = k_1 + ik_2$$

where k_1 and k_2 are real and in \mathbf{K}_L , since \mathbf{K}_L is a linear manifold (see 10.42). Let

$$[v_r, k_r] \in L(p),$$

$r = 1, 2$. Then we have (P2)

$$[v_1 + iv_2, k] \in L(p).$$

Then

$$J_k(p) = (v_1, k_1) + (v_2, k_2) + i(v_1, k_2) - i(v_2, k_1).$$

Now by P5(I), $(v_1, k_2) = (v_2, k_1)$. Hence

$$J_k(p) = (v_1, k_1) + (v_2, k_2) \quad (3)$$

for any $p \in \Gamma_L$. Since each summand in (3) is a PR function, it follows that $J_k(p)$ is PR for any $k \in \mathbf{K}_L$.

12.12 Let \mathbf{K}_1 be the set of all vectors $k \in \mathbf{K}_L$ such that

$$J_k(p) = 0 \quad \text{for every } p \in \Gamma_L.$$

Notice that we do not assert that \mathbf{K}_1 is a linear manifold.

If $k \in \mathbf{K}_1$ then $k \in \mathbf{K}_L$ and (3) above applies. Then

$$(v_1, k_1) + (v_2, k_2) = 0$$

and, using this and the PR property of each summand, we conclude that k_1 and k_2 are in \mathbf{K}_1 .

12.13 We wish now to show that $\mathbf{K}_1 \subseteq \mathbf{K}_{L_0}(p)$. Consider a real $j \in \mathbf{K}_L$ and a real $k \in \mathbf{K}_1$. Let

$$\begin{aligned} [u(p), j] \in L(p), \\ [v(p), k] \in L(p). \end{aligned} \quad (4)$$

Then, given any real λ , by P2

$$[u(p) + \lambda v(p), j + \lambda k] \in L(p).$$

Then, because $k \in \mathbf{K}_1$,

$$(u + \lambda v, j + \lambda k) = (u, j) + \lambda(v, j) + \lambda(u, k).$$

Since j and k are real, by P5(I) this can be written

$$(u + \lambda v, j + \lambda k) = (u, j) + 2\lambda(v, j). \quad (5)$$

Choose any p_1 such that $\text{Re}(p_1) \geq 0$. Then P7(I) implies that the left side of (5) has a non-negative real part at $p = p_1$. The right side, by suitable choice of λ , can have any chosen real part unless

$$\text{Re}(v(p_1), j) = 0. \quad (6)$$

Hence P7(I) implies (6). Now $(v(p), j)$ is a rational function, by P6(I) applied to the other members of (5). By (6), this rational function has a vanishing real part throughout the right half p -plane. Hence it is an imaginary constant:

$$(v(p), j) \equiv ia. \quad (7)$$

Then

$$\overline{(v(p), j)} = \overline{(v(p), j)} = -ia. \quad (8)$$

But $[v(p), k] \in L(p)$, so $\overline{[v(p), k]} \in L(\bar{p})$ by P4. Since also $[v(\bar{p}), k] \in L(\bar{p})$, by 12.1, we have from (8) that

$$(v(\bar{p}), j) = -ia.$$

Comparing this with (7) written for \bar{p} , we have $a = 0$ and

$$(v(p), j) = 0 \quad \text{for } p \in \Gamma_L. \quad (9)$$

Now $v(p)$ was determined by (4) wherein k is real. For any $k \in \mathbf{K}_1$, $k = k_1 + ik_2$, where k_1 and k_2 are real and in \mathbf{K}_1 (12.11). A corresponding $v(p)$ satisfying (4) can be written

$$v(p) = v_1(p) + iv_2(p), \quad (10)$$

by P2, where $[v_r(p), k_r] \epsilon L(p)$, $r = 1, 2$. Then (9) holds for each of $v_1(p)$, $v_2(p)$ and therefore also for the $v(p)$ of (10).

We have showed now that for any $p \epsilon \Gamma_L$ and any $k \epsilon \mathbf{K}_1$, the $v(p)$ of (4) has a vanishing scalar product with every real $j \epsilon \mathbf{K}_L$. Since \mathbf{K}_L is spanned by real j ,

$$v(p) \epsilon (\mathbf{K}_L)^0 = \mathbf{V}_{L0}. \quad (11)$$

12.14 By (11),

$$[v(p), 0] \epsilon L(p).$$

Comparing this with (4), and applying P2,

$$[v(p) - v(p), k - 0] = [0, k] \epsilon L(p).$$

Since k is now any vector in \mathbf{K}_1 , we have

$$\mathbf{K}_1 \subseteq \mathbf{K}_{L0}(p) \subseteq \mathbf{K}_L \quad (12)$$

for every $p \epsilon \Gamma_L$.

12.15 We can now also show that $\mathbf{V}_L(p) \subseteq (\mathbf{K}_1)^0$. We return to 12.13 and read (9) thereof as originally derived for real j and k . Applying P5(I), we have from (9) that

$$(u(p), k) = 0 \quad \text{for } p \epsilon \Gamma_L. \quad (13)$$

By the argument immediately following (9), (13) also holds for any $k \epsilon \mathbf{K}_1$, provided j is real. As in 12.11 any $j \epsilon \mathbf{K}_L$ can be written $j = j_1 + ij_2$, where j_1 and j_2 are real, and the corresponding

$$u(p) = u_1(p) + iu_2(p)$$

where $[u_r(p), j_r] \epsilon L(p)$. Therefore, finally, (13) holds for any $u(p)$ satisfying (4)—i.e., any $u(p) \epsilon \mathbf{V}_L(p)$ —and any $k \epsilon \mathbf{K}_1$. Therefore

$$\mathbf{V}_L(p) \subseteq (\mathbf{K}_1)^0 \quad (14)$$

for any $p \epsilon \Gamma_L$.

12.2 We now fix our attention on a specific *real* $p_0 \epsilon \Gamma_L$

12.21 By P4, if

$$[v, k] \epsilon L(p_0)$$

we have also

$$[\bar{v}, \bar{k}] \epsilon L(\bar{p}_0) = L(p_0).$$

In particular, $\mathbf{K}_{L0}(p_0)$ is real.

12.22 We can now show that \mathbf{K}_1 is a real linear manifold. Consider a real $k \in \mathbf{K}_{L^0}(p_0)$. Then $[0, k] \in L(p_0)$ and by 12.1

$$J_k(p_0) = 0.$$

Then by 12.11 (and 15.12), $J_k(p) = 0$, so $k \in \mathbf{K}_1$. Since $\mathbf{K}_{L^0}(p_0)$ is spanned by real k (12.21), we have

$$\mathbf{K}_{L^0}(p_0) \subseteq \mathbf{K}_1.$$

Comparing this with (12) gives us

$$\mathbf{K}_{L^0}(p_0) = \mathbf{K}_1. \quad (15)$$

Since $\mathbf{K}_{L^0}(p_0)$ is a real linear manifold by definition and 12.21, we see that \mathbf{K}_1 is.

12.3 Let us now write, by (12) and (15),

$$\mathbf{K}_L = \mathbf{K}_2 \oplus \mathbf{K}_1 \quad (16)$$

where \mathbf{K}_2 is an arbitrary fixed manifold disjoint from \mathbf{K}_1 and with it spanning \mathbf{K}_L . All three manifolds are real (12.21, (15), 10.6).

Choose a \mathbf{K}_3 disjoint from \mathbf{K}_L such that

$$\mathbf{K} = \mathbf{K}_3 \oplus \mathbf{K}_2 \oplus \mathbf{K}_1. \quad (17)$$

Let the decomposition of \mathbf{V} dual to (17) be (10.6)

$$\mathbf{V} = \mathbf{V}_3 \oplus \mathbf{V}_2 \oplus \mathbf{V}_1.$$

Then $\mathbf{V}_3 = (\mathbf{K}_2 \oplus \mathbf{K}_1)^0 = (\mathbf{K}_L)^0 = \mathbf{V}_{L^0}$ by 12.01. Hence

$$\mathbf{V} = \mathbf{V}_{L^0} \oplus \mathbf{V}_2 \oplus \mathbf{V}_1. \quad (18)$$

By (14) and the definitions,

$$\mathbf{V}_{L^0} \subseteq \mathbf{V}_L(p) \subseteq \mathbf{V}_{L^0} \oplus \mathbf{V}_2. \quad (19)$$

12.31 Consider a real p_0 . Then by P3'(I), (15) and (16) we have

$$\mathbf{K}_{L^0}(p_0) \subseteq \mathbf{K}_L(p_0) \subseteq \mathbf{K}_2 \oplus \mathbf{K}_{L^0}(p_0). \quad (20)$$

This is now an expression dual to (19). We shall prove next that, given any $k \in \mathbf{K}_L(p_0) \cap \mathbf{K}_2 (= \mathbf{K}_2)$, there is a unique $v_k \in \mathbf{V}_L(p_0) \cap \mathbf{V}_2$ such that

$$[v_k, k] \in L(p_0). \quad (21)$$

Dually, given any $v \in \mathbf{V}_L(p_0) \cap \mathbf{V}_2$, there is a unique $k_v \in \mathbf{K}_L(p_0) \cap \mathbf{K}_2$ such that

$$[v, k_v] \in L(p_0).$$

The proof is a standard one in algebra and depends only upon P2, (19), and (20).

Proof: Given $k \in \mathbf{K}_L(p_0) \cap \mathbf{K}_2$, there is some $v \in \mathbf{V}_L(p_0)$ such that

$$[v, k] \in L(p_0). \quad (22)$$

By (19), then,

$$v = v_0 + v_2$$

where $v_0 \in \mathbf{V}_{L0}$, $v_2 \in \mathbf{V}_2$. Then

$$[v_0, 0] \in L(p_0)$$

so, applying P2 to this and (22),

$$[v - v_0, k - 0] = [v_2, k] \in L(p_0). \quad (23)$$

Hence $v_2 \in \mathbf{V}_L(p_0) \cap \mathbf{V}_2$ and $v_k = v_2$ satisfies (21). Suppose now $v_3 \in \mathbf{V}_L(p_0) \cap \mathbf{V}_2$ and

$$[v_3, k] \in L(p_0).$$

Then using this with (23) and P2

$$[v_2 - v_3, 0] \in L(p_0).$$

Hence $(v_2 - v_3) \in \mathbf{V}_{L0}$. Now $\mathbf{V}_L(p_0) \cap \mathbf{V}_2$ is a linear manifold and contains v_2, v_3 . Hence

$$(v_2 - v_3) \in \mathbf{V}_{L0} \cap \mathbf{V}_L(p_0) \cap \mathbf{V}_2 = 0.$$

Therefore $v_2 = v_3$.

The dual argument completes the proof.

12.32 The argument actually exhibited in 12.31 uses only P2 and (19), hence the v_k of (21) is unique whether or not p_0 is real. Indeed, this is true even when $k \in \mathbf{K}_L$.

12.33 The result of 12.31 establishes a bi-unique linear mapping between \mathbf{K}_2 and $\mathbf{V}_L(p_0) \cap \mathbf{V}_2$. Hence these two manifolds are of the same dimension. Since \mathbf{K}_2 and $\mathbf{V}_2 = \mathbf{K}_2^*$ are of the same dimension by construction, it follows that

$$\mathbf{V}_L(p_0) \cap \mathbf{V}_2 = \mathbf{V}_2$$

and, by (19), that

$$\mathbf{V}_L(p_0) = \mathbf{V}_{L0} \oplus \mathbf{V}_2.$$

12.4 Let us now introduce a real frame in \mathbf{V} and \mathbf{K} which provides real bases in $\mathbf{K}_1, \mathbf{K}_2, \mathbf{K}_3$ and in $\mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_{L0}$ of (17) and (18). Let k_1, \dots, k_m

be the basis vectors spanning \mathbf{K}_2 . By 12.32, there are unique vectors $u_1(p), \dots, u_m(p)$ in \mathbf{V}_2 such that

$$[u_r(p), k_r] \epsilon L(p).$$

Let v_1, \dots, v_m be the (real) basis vectors in \mathbf{V}_2 dual to the k_1, \dots, k_m :

$$(v_r, k_s) = \delta_{rs} \quad 1 \leq r \leq s. \quad (24)$$

Since the $u_r(p)$ are all in \mathbf{V}_2 we have for each $p \in \Gamma_L$

$$u_s(p) = \sum_{r=1}^m a_{rs}(p) v_r \quad (25)$$

where the coefficients $a_{rs}(p)$ are calculated by (24) to be

$$a_{rs}(p) = (u_s(p), k_r). \quad (26)$$

12.41 Because the k_r are real, P5(I) implies that

$$a_{sr}(p) = (u_r(p), k_s) = (u_s(p), k_r) = a_{rs}(p). \quad (27)$$

By the reasoning just following (8) and by the uniqueness of the $u_s(p) \epsilon \mathbf{V}_2$, since \mathbf{V}_2 is real, we have $\overline{u_s(\bar{p})} = u_s(\bar{p})$. Then

$$\overline{a_{rs}(p)} = \overline{(u_s(p), k_r)} = (u_s(\bar{p}), k_r) = a_{rs}(\bar{p}).$$

12.42 We have by P2 that

$$[u_r(p) + \lambda u_s(p), k_r + \lambda k_s] \epsilon L(p), \quad (28)$$

for any λ . The identity

$$\begin{aligned} (u_r + u_s, k_r + k_s) - (u_r - u_s, k_r - k_s) \\ = 2(u_r, k_s) + 2(u_s, k_r) \end{aligned} \quad (29)$$

holds in fact for any vectors u_r, u_s, k_r, k_s . Using (27), (28) and P6(I), it exhibits $a_{rs}(p)$ as a rational function.

12.5 Consider the $m \times m$ matrix $[Z_1(p)]$ whose elements are the $a_{rs}(p)$. the s -th column of this matrix consists of the components of $u_s(p)$. The rank of the matrix is by definition the dimension of the space spanned by these columns.

12.51 Now the rank of $[Z_1(p)]$ can be expressed in terms of the vanishing or not of its various minor determinants. There are finitely many such minors and each is a rational function. Each is either identically zero or else vanishes at only finitely many points. Hence the rank of $[Z_1(p)]$, except at these finitely many points, and at the p in the comple-

ment of Γ_L , is a constant. We call this constant the nominal rank of $[Z_1(p)]$.

12.52 Let Γ'_L consist of all $p \in \Gamma_L$ where $[Z_1(p)]$ has its nominal rank. Then Γ'_L has a finite complement. By the reality result of 12.41, if $p \in \Gamma'_L$ then $\bar{p} \in \Gamma'_L$.

It is clear that at any $p \in \Gamma_L$ the rank of $[Z_1(p)]$ does not exceed its nominal rank.

12.53 By construction, the vectors $u_1(p), \dots, u_m(p)$ all lie in $\mathbf{V}_L(p) \cap \mathbf{V}_2$. By the reasoning of 12.33, at any real $p_0 \in \Gamma_L$ they span \mathbf{V}_2 . Hence the nominal rank of $[Z_1(p)]$ is m . Therefore, for any $p \in \Gamma'_L$, $[Z_1(p)]$ has rank m and the $u_1(p), \dots, u_m(p)$, lying in \mathbf{V}_2 , still span \mathbf{V}_2 . Therefore for all $p \in \Gamma'_L$

$$\mathbf{V}_L(p) \cap \mathbf{V}_2 = \mathbf{V}_2.$$

By (19), then,

$$\mathbf{V}_L(p) = \mathbf{V}_{L0} \oplus \mathbf{V}_2 = \mathbf{V}_L, \quad (30)$$

a fixed manifold, for all $p \in \Gamma'_L$.

12.54 It is clear by its construction (cf. Halmos⁹, par. 26) that $[Z_1(p)]$ describes the mapping of 12.32 from \mathbf{K}_2 to $\mathbf{V}_2 = \mathbf{V}_L(p) \cap \mathbf{V}_2$ by

$$[v_k] = [Z_1(p)][k].$$

Here the m -tuples $[v_k]$ and $[k]$ are the components of v_k and k relative to the bases now available in \mathbf{V}_2 and \mathbf{K}_2 .

12.55 We repeat

$$\mathbf{K}_1 \subseteq \mathbf{K}_{L0}(p) \subseteq \mathbf{K}_L = \mathbf{K}_1 \oplus \mathbf{K}_2. \quad (12)$$

Fix a $p \in \Gamma'_L$ and a $k \in \mathbf{K}_{L0}(p) \cap \mathbf{K}_2$. Then $[0, k] \in L(p)$. Since $0 \in \mathbf{V}_2$, it follows from 12.54 that $[Z_1(p)]$ annihilates k . Suppose $m \neq 0$. Since the rank of $[Z_1(p)]$ is m , it follows that $k = 0$. Hence for $p \in \Gamma'_L$

$$\mathbf{K}_{L0}(p) \cap \mathbf{K}_2 = 0.$$

By (12), then,

$$\mathbf{K}_{L0}(p) = \mathbf{K}_1 = \mathbf{K}_{L0}, \quad (31)$$

a fixed manifold. This, with the result of 12.53, proves that L satisfies P3(A), when $m \neq 0$.

If $m = 0$ then $\mathbf{V}_2 = 0$, $\mathbf{K}_2 = 0$ and (31) follows from (12) and (16).

12.56 $[Z_1(p)]$ is of dimension m and rank m for any $p \in \Gamma'_L$. Therefore

the correspondence of 12.32 and 12.54 between \mathbf{V}_2 and \mathbf{K}_2 is bi-unique for any $p \in \Gamma'_L$. This extends 12.31 to any $p \in \Gamma'_L$.

12.57 If $m = 0$, i.e., if $\mathbf{V}_2 = \mathbf{K}_2 = 0$, then $\mathbf{V}_{L0} = (\mathbf{K}_{L0})'$ and the fact that L satisfies all the postulates is trivial because all scalar products (v, k) for $v \in \mathbf{V}_L = \mathbf{V}_{L0}$ and $k \in \mathbf{K}_L = \mathbf{K}_{L0}$ are zero. If $m \neq 0$, we have yet to show that L satisfies P5(A), P6(A), P7(A).

12.6 Since now L satisfies P3, 7.7 as given is applicable and we find (with 12.1) that if $p \in \Gamma'_L$ and

$$[v, k] \in L(p),$$

then (v, k) is fixed by either v or k . Furthermore,

$$(v, k) = (v + v_0, k + k_0)$$

for any $v_0 \in \mathbf{V}_{L0}$, $k_0 \in \mathbf{K}_{L0}$.

12.61 If $p \in \Gamma'_L$ and $[v, k] \in L(p)$, then $v \in \mathbf{V}_L$, $k \in \mathbf{K}_L$. By (30), (31), and (16), therefore, there exist $v_0 \in \mathbf{V}_{L0}$, $k_0 \in \mathbf{K}_{L0}$ such that $u = v - v_0 \in \mathbf{V}_2$, $j = (k - k_0) \in \mathbf{K}_2$. Then by P2

$$[u, j] \in L(p). \quad (32)$$

By 12.6, then, any value assumed by a scalar product (v, k) with $[v, k] \in L(p)$ is also assumed by a product (u, j) , where (32) holds and $u \in \mathbf{V}_2$, $j \in \mathbf{K}_2$.

XIII. SUFFICIENCY OF THE POSTULATES

13.0 We suppose that L satisfies the postulates of 12.0. Then the results of Section 12 are applicable. The ones of first importance are contained in the facts from (15), (30) and (31), that

$$\mathbf{V}_L = \mathbf{V}_{L0} \oplus \mathbf{V}_2,$$

$$\mathbf{K}_L = \mathbf{K}_2 \oplus \mathbf{K}_{L0},$$

where the choice of \mathbf{K}_2 was governed only by the requirement that the second of these formulae hold.

13.01 Considering \mathbf{K}_2 and \mathbf{V}_2 as separate spaces, $\mathbf{V}_2 = \mathbf{K}_2^*$ by 10.6. Let M be the geometrical linear correspondence between them with frequency domain Γ'_L and pairs described by 12.31 and 12.56 (or 12.54). That is, as vectors in \mathbf{V}_2 and \mathbf{K}_2

$$[v, k] \in M(p)$$

if and only if, as vectors in \mathbf{V} and \mathbf{K} ,

$$[v, k] \in L(p).$$

13.02 In the real frame of 12.4 let us renumber the basis vectors so that

$$\begin{aligned} v_1, \dots, v_r & \text{ span } \mathbf{V}_{L0}, \\ v_{r+1}, \dots, v_{r+m} & \text{ span } \mathbf{V}_2, \\ v_{r+m+1}, \dots, v_n & \text{ span } \mathbf{V}_1. \end{aligned}$$

Then

$$\begin{aligned} k_1, \dots, k_{r+m} & \text{ span } \mathbf{K}_2, \\ k_{r+m+1}, \dots, k_n & \text{ span } \mathbf{K}_{L0}. \end{aligned}$$

We say that such a frame *reduces* L .

13.1 Let us now interpret the s -th components of $[v]$ and $[k]$ in this frame respectively as the voltage across and the current in an ideal branch β_s of a $2n$ -pole \mathbf{N} , $1 \leq s \leq n$.

By construction, the vectors $v \in \mathbf{V}_L$ in this frame have components $a_{r+m+1} = \dots = a_n = 0$, since v_1, \dots, v_{r+m} span \mathbf{V}_L . At the same time, the components b_{r+m+1}, \dots, b_n of $[k]$ may be chosen arbitrarily without altering the fact that $[[v], [k]] \in [L](p)$ because of 12.06. Therefore, the ideal branches $\beta_{r+m+1}, \dots, \beta_n$ can each be realized physically by a short circuit.

In a dual way, since k_{r+1}, \dots, k_n span \mathbf{K}_L , any $k \in \mathbf{K}_L$ has components b_1, \dots, b_r all zero in our chosen frame. Furthermore, the components a_1, \dots, a_r of $[v]$ can be chosen at will. Hence the ideal branches β_1, \dots, β_r can each be realized physically by an open circuit.

Let \mathbf{N}_1 now be the $2m$ -pole whose ideal branches are $\beta_{r+1}, \dots, \beta_{r+m}$. Let the pairs $[[v], [k]]$ admitted by \mathbf{N}_1 at each $p \in \Gamma'_L$ be the $[[v], [k]]$, where $[v, k] \in M(p)$ (13.01). The representation just found for \mathbf{N} shows that \mathbf{N} is physically realizable if and only if \mathbf{N}_1 is.

13.11 The matrix $[Z_1(p)]$ of 12.54 is the impedance matrix of the $2m$ -pole \mathbf{N}_1 .

13.12 We now show that $[Z_1(p)]$ is a positive real matrix. The displayed formulae of 12.41 show (ii) and (iii) of 1.1, and 12.42 shows (i). Now suppose that $[v, k] \in M(p)$. Then, as vectors in \mathbf{V} and \mathbf{K} , $[v, k] \in L(p)$ by definition of $M(p)$. Then, however, if k is fixed

$$J_k(p) = (v, k)$$

is a PR function (12.11). Regarding v and k in \mathbf{V}_2 and \mathbf{K}_2 let

$$[b_{r+1}, \dots, b_{r+m}] = [k].$$

Then by (1) of 7.0

$$(v, k) = \sum_{t=1}^m \sum_{s=1}^m a_{st}(p) b_{t+r} \bar{b}_{s+r}$$

and this has a non-negative real part if $\text{Re}(p) > 0$. This is (iv) of 1.1.

13.2 We can now prove the lemmas 8.1 and 8.2. Given a linear correspondence $[L]$ which satisfies P1, \dots , P7 by 11.13 we can interpret $[L]$ as the concrete correspondence representing a geometrical correspondence L in some chosen real frame, and L satisfies P1, \dots , P7. Then by the results in 13.01–13.12 there exists a real frame in which the representative $[L]_1$ of L has the properties claimed in 8.1 and 8.2 for L_w . But we saw in 11.22 that $[L]$ and $[L]_1$ are related by a real matrix W like the L and L_w of Section 8. Q.E.D.

13.21 With the proofs of 8.1 and 8.2 we have reduced the sufficiency claimed for P1, \dots , P7 in 8.0 to the sufficiency of positive reality of $[Z(p)]$ claimed in 1.1, by the argument outlined in 8.5.

XIV. OPERATOR-VALUED FUNCTIONS OF p

The next three sections are directed principally toward the proof of the matrix theorem of 1.1. They do however, contribute to 12.10 and to the necessity proof.

14.0 We continue to use the geometric language. The reader who regards this as unduly pedantic is free to place a concrete interpretation upon every argument, for all of the arguments are either frankly based on matrix representations or upon the three identities:

$$14.01 \quad (Zj, k) = \overline{(Z^*k, j)} \text{ for all } j, k \in \mathbf{K}.$$

$$14.02 \quad \bar{Z}\bar{k} = \overline{(Zk)} \text{ for all } k \in \mathbf{K}.$$

$$14.03 \quad Z' = (\bar{Z})^* = \overline{(Z^*)}$$

14.04 These identities are obvious for matrices using 7.0 and 7.2. Geometrically, the first and second define Z^* and \bar{Z} , and the third defines Z' in two ways. The equivalence of these two ways is a theorem based on (10) of 10.33.

14.05 The symbol Z will always denote an impedance (operator, matrix, scalar), and Y will always denote an admittance. An impedance oper-

ates from \mathbf{K} to \mathbf{V} , an admittance dually. The operators in Halmos⁹ are physically dimensionless, in that they operate, e.g., from \mathbf{V} to \mathbf{V} . This difference is scarcely noticeable.

We shall regularly omit the duals to concepts or proofs given in terms of impedances. In doing so, we adopt the rule that the dual to an expression

$$(Zk, k)$$

is

$$\overline{(v, Yv)}.$$

14.1 An operator is called symmetric if $Z = Z'$. Such operators have three useful special properties:

14.11 If Z is symmetric and j and k are real, then

$$(Zj, k) = \overline{(\bar{Z}j, k)} = ((\bar{Z})^*k, j) = (Z'k, j) = (Zk, j)$$

by (10) of 10.33, 14.02, 14.01, 14.03, and hypothesis.

14.12 Let $k = k_1 + ik_2$, where k_1 and k_2 are real (10.42). If Z is symmetric then

$$(Zk, k) = (Zk_1, k_1) + (Zk_2, k_2),$$

for, by 14.11,

$$\begin{aligned} (Zk_1, ik_2) &= -i(Zk_1, k_2) = -i(Zk_2, k_1) \\ &= -(Z(ik_2), k_1). \end{aligned}$$

(Cf. the similar identity in 12.11.)

14.13 The symmetric operator Z is completely defined by the quadratic form

$$(Zk, k) \tag{1}$$

as a function of *real* $k \in \mathbf{K}$. For 14.11 permits the formula (29) of 12.42 in any real frame, where $u_s = Zk_s$. The matrix elements of $[Z(p)]$ in that frame are then defined by that formula in terms of values of (1) for real k .

The form (1) specifies any Z (symmetric or not) if k is allowed to range over all of \mathbf{K} (Halmos⁹, par. 53).

14.2 Let $Z(p)$ now be an impedance operator depending on p . We say that $p_0 \neq \infty$ is a pole of order m of $Z(p)$ if

$$\ell(k) = \lim_{p \rightarrow p_0} (p - p_0)^m (Z(p)k, k) \tag{2}$$

exists for every $k \in \mathbf{K}$ and is not identically zero. By 15.13, this limit $\ell(k)$ defines an operator R_0 , the residue* of $Z(p)$ at p_0 , by

$$(R_0 k, k) = \ell(k) \quad \text{for } k \in \mathbf{K}.$$

The changes in (2) required to define a pole at $p = \infty$ are obvious.

14.21 A pole p_0 of order m of $Z(p)$ is a pole of some matrix element of $[Z(p)]$, of order m , in any frame, and no element of $[Z(p)]$ has a pole at p_0 of order exceeding m . For the elements of $[Z(p)]$ are defined by the values of $(Z(p)k, k)$, by 14.11 and Halmos⁹ loc. cit.

XV. POSITIVE REAL FUNCTIONS

15.0 Let $f(p)$ be a scalar function of the complex variable p . Following Brune² we define $f(p)$ to be positive real if

- (i) $f(p)$ is a rational function of p ,
- (ii) $\overline{f(p)} = f(\bar{p})$,
- (iii) $\operatorname{Re}(p) > 0$ implies $\operatorname{Re}(f(p)) \geq 0$.

The property (i) of being rational is of course on a quite different level of ideas from the other properties, but it saves words later to include it specifically in the meaning of positive real.

We abbreviate the words positive real to PR.

15.01 The open region of the complex plane consisting of all finite p such that $\operatorname{Re}(p) > 0$ —the right half plane—we denote by Γ_+ .

15.1 Brune, loc. cit., established a number of properties of PR functions $f(p)$ which will be useful to us here:

15.11 $f(p)$ has no poles in Γ_+ .

15.12 If $\operatorname{Re}(f(p)) = 0$ for some $p \in \Gamma_+$, then $f(p) \equiv 0$ for all p .

15.13 If it exists, $\frac{1}{f(p)}$ is PR.

15.14 If $f(p)$ has a pole at $p = p_0$, it has one at $p = \bar{p}_0$.

15.15 If $f(p)$ has a pole at $p = i\omega_0$, that pole is simple and

$$f(p) = \frac{2p}{p^2 + \omega_0^2} r + f_1(p),$$

where $r > 0$, and $f_1(p)$ is PR.

* Properly, R_0 is a residue only when $m = 1$. There is no convenient name available for general m .

15.16 If $f(p)$ has a pole at $p = \infty$, that pole is simple and

$$f(p) = pr + f_1(p),$$

where $r > 0$, and $f_1(p)$ is PR.

15.17 We shall use all of these in the next section, save 15.13. Our aim is to prove properties analogous to 15.11, \dots , 15.16 for PR matrices and operators.

The reader familiar with the Brune process² for realization of a 2-pole will remember the importance of the properties 15.11, \dots , 15.16 for the success of that process. Correspondingly, we must establish the analogs of these properties to implement the general Brune process for $2n$ -poles.

XVI. POSITIVE REAL OPERATORS

16.0 An operator $Z(p)$ from \mathbf{K} to \mathbf{V} will be called positive real (PR) if in some real coordinate frame the matrix $[Z(p)]$ is a PR matrix in the sense of 1.1—that is

(i) $[Z(p)]$ has rational elements $Z_{rs}(p)$

(ii) $\overline{Z_{rs}(p)} = Z_{rs}(\bar{p})$

(iii) $Z_{rs}(p) = Z_{sr}(p)$

(iv) For any real $k \in \mathbf{K}$ and any $p \in \Gamma_+$

$$\operatorname{Re}(Z(p)k, k) \geq 0.$$

We intend in this section to establish for PR operators the properties listed below. By subtracting 0.9 from the designation of each property one obtains the designation of the analogous property of a PR scalar function, stated earlier.

16.01 $Z(p)$ has no poles in Γ_+ .

16.02 If $\operatorname{Re}(Z(p)k, k) = 0$ for some $p \in \Gamma_+$, then $Z(p)k \equiv 0$ for all p .

16.03 If it exists, $Z^{-1}(p) = Y(p)$ is PR.

16.04 If $Z(p)$ has a pole at $p = p_0$, it has one at $p = \bar{p}_0$.

16.05 If $Z(p)$ has a pole at $p = i\omega_0$, that pole is simple* and

$$Z(p) = \frac{2p}{p^2 + \omega_0^2} R + Z_1(p),$$

where R is real, symmetric, and semi-definite, not zero, and $Z_1(p)$ is PR.

* i.e., of order one.

16.06 If $Z(p)$ has a pole at $p = \infty$, that pole is simple and

$$Z(p) = pR + Z_1(p)$$

where $R = R' = \bar{R}$, $R \geq 0$ and $Z_1(p)$ is PR.

16.07 There is property of rational scalar functions $f(p)$, whether PR or not, that is essential in the Brune theory: the existence of a finite integer, the degree of f . Each step in the Brune reduction of $f(p)$ leaves an unreduced portion which is of lower degree than the function upon which the step was performed. The finiteness of the original degree of f then guarantees the termination of the process in finitely many steps.

There exists also for rational matrices (and operators) a concept of degree. This degree plays the same role in the general Brune process for $2n$ -poles as the degree of a scalar function does in the process for 2-poles. To define this degree and develop its properties requires an excursion into classical algebra. Since we shall not need these ideas until Part II we defer further discussion of them to that part.

16.1 If $Z(p)$ is PR it follows at once that the matrix $[Z(p)]$ is PR in any real frame.

Proof: Two such matrices are related by

$$[Z(p)]_1 = [U][Z(p)][U']$$

where U is real, by 11.22 and the argument in 8.6. The PR properties of $[Z(p)]$ are obviously preserved by this operation.

16.11 If $Z(p)$ is PR, then

$$Z(p) = Z'(p) = \overline{Z^*(p)} = \overline{Z(\bar{p})}.$$

Proof: Use 16.0 and 14.03 in a real frame.

16.12 If $Z(p)$ is PR, then for any given $k \in \mathbf{K}$ the function

$$J_k(p) = (Z(p)k, k)$$

is a PR scalar function. It follows that the limitation in (iv) of 16.0 to real k is a simplification, not a restriction.

Proof: $J_k(p)$ is independent of coordinate representation. By use of a real frame, (i) of 16.0 implies (i) of 15.0.

By 14.01 and 16.11

$$\overline{J_k(p)} = (Z^*(p)k, k) = (Z(\bar{p})k, k) = J_k(\bar{p}).$$

This is (ii) of 15.0. For any k , 14.12 and (iv) of 16.0 imply (iii) of 15.0.

16.13 Conversely to 16.12, if $Z(p)$ is symmetric and $J_k(p)$ is PR for every real k , then $Z(p)$ is PR, and $J_k(p)$ is PR for all k .

Proof: $J_k(p)$ is rational so (i) of 16.0 holds in any frame by 14.13. Clearly (iv) of 16.0 holds.

Now for real k , by (10) of 10.33 and 14.02

$$J_k(\bar{p}) = \overline{J_k(p)} = \overline{(Z(p)k, k)}.$$

Hence $Z(\bar{p}) = \overline{Z(p)}$ by 14.13. This is (ii) of 16.0, and (iii) there holds by hypothesis.

16.2 *Proof of 16.01:* By 15.11 and 16.12, $J_k(p)$ has no poles in Γ_+ . This is 16.01 by the definition 14.3 of pole.

16.21 *Corollary:* Any PR $Z(p)$ can be considered as defined throughout Γ_+ : for any k , $J_k(p)$ is defined throughout Γ_+ by 16.2. For each p , as a function of k , $J_k(p)$ defines $Z(p)$ (14.13).

16.3 *Proof of 16.03:* In any frame $[Z^{-1}(p)] = [Z(p)]^{-1} = [Y(p)]$ consists of rational elements, by direct calculation of the inverse matrix. In a real frame $[Y(p)] = [Z^{-1}(p)]$ is symmetric and real for real p by the same argument (both facts are also deducible geometrically). Hence we have the duals of (i), (ii) and (iii) of 16.0 for $Y(p)$. Clearly $Y(p)$ is defined throughout Γ_+ .

Now suppose that for some $v \in \mathbf{V}$ and some $p_0 \in \Gamma_+$ we have

$$\operatorname{Re}(\overline{v, Y(p_0)v}) < 0.$$

Then there is a $k \in \mathbf{K}$ such that $v = Z(p_0)k$. Therefore

$$\operatorname{Re}(\overline{Z(p_0)k, k}) = \operatorname{Re}(Z(p_0)k, k) < 0.$$

Since this is impossible, we have the dual of (iv) of 16.0 for $Y(p)$ and $Y(p)$ is PR.

16.4 *Proof of 16.04:* This is immediate from 15.14, 14.3, and 16.12.

16.5 *Proofs of 16.05 and 16.06:* Suppose $Z(p)$ has a pole at $p = i\omega_0$. Then $(Z(p)k, k)$ does and that pole is simple by 15.15 and 16.12. Then by 14.3 we can write

$$Z(p) = \frac{1}{p - i\omega_0} R_0 + Z_0(p)$$

where $Z_0(p)$ is regular at $p = i\omega_0$. Now $Z_0(p)$ has a pole at $p = -i\omega_0$ by 16.5, so a similar argument gives

$$Z(p) = \frac{1}{p - i\omega_0} R_0 + \frac{1}{p + i\omega_0} R_1 + Z_1(p), \quad (1)$$

where $Z_1(p)$ has no pole at $i\omega_0$ or $-i\omega_0$. The symmetry of Z and linear independence of the terms above then imply the symmetry of R_0 , R_1 and $Z_1(p)$.

For any $k \in \mathbf{K}$, now,

$$(Z(p)k, k) = \frac{1}{p - i\omega_0} (R_0 k, k) + \frac{1}{p + i\omega_0} (R_1 k, k) + (Z_1(p)k, k).$$

Applying 16.12 and 15.15,

$$(R_0 k, k) = (R_1 k, k) \geq 0$$

for all k . Hence $R_0 = R_1 = R$ (say) and R is semi-definite. Also, $(Z_1(p)k, k)$ appears as the residue $f_1(p)$ in 15.15 and is therefore PR. Then $Z_1(p)$ is PR by 16.13. With R_0 and R_1 identified, (1) above is the expansion given in 16.05. We have now proved all of 16.05 save the reality of R . But

$$\frac{2p}{p^2 + \omega_0^2} R$$

is PR, by 16.13, hence is real for real p . Therefore R is real.

The proof of 16.06 is similar.

16.6 To prove 16.02 we appear to digress somewhat, by first completing the proof of the fundamental lemma of 12.0. It was established in Section 13 that the matrix $[Z_1(p)]$ describing $M(p)$ in the chosen basis is PR. The case in which it is nonsingular (i.e., $m \neq 0$, cf. 12.56, 12.57) remains to be examined.

16.61 If $[Z_1(p)]$ is nonsingular then its inverse is PR (16.3). Then for any $v \in \mathbf{V}_2$,

$$\overline{(v, k)} = \overline{(v, Y(p)v)} \quad (2)$$

is PR (16.12 dual). By 12.61, for any $u \in \mathbf{V}_L$, the values of the function $F_u(p)$ are the values of (2) for some $v \in \mathbf{V}_2$. Hence $F_u(p)$ is PR. This is P6(A) and P7(A) for L .

16.62 To settle P5 for L in 12.0, consider $p \in \Gamma'_L$ and

$$[v, k] \in L(p), \quad [u, j] \in L(p),$$

where u and v are real. Then, say,

$$v = v_0 + v_1,$$

where $v_0 \in V_{L0}$, $v_1 \in V_{L2}$. But then

$$v = \bar{v} = \bar{v}_0 + \bar{v}_1$$

and, because \mathbf{V}_{L0} and \mathbf{V}_2 are real, $\bar{v}_0 = v_0$, $\bar{v}_1 = v_1$, and these vectors are real. Using similar reasoning for u ,

$$(v, j) = (v_1, Y(p)u_1), \quad (u, k) = (u_1, Y(p)v_1), \quad (3)$$

by 12.61. The equality $(u, k) = (v, j)$ now follows from (3) and the duals of 16.11, 14.11. Hence we have P5(A) for L and 12.0 is proved.

16.7 We now prove an important

Lemma: Let $Z(p)$ be a PR operator from \mathbf{K} to \mathbf{V} . Let Γ_L be the set of p where $Z(p)$ is defined and has a rank equal to its nominal rank. Let L be the correspondence with domain Γ_L and pairs

$$[Z(p)k, k], \quad k \in \mathbf{K}_L.$$

Then L satisfies P1, \dots , P7.

Proof: L satisfies P1 and P2 (6.3). Γ_L satisfies P4 by the argument of 12.52. Then L satisfies P4, for by 16.11

$$\overline{Z(p)k} = Z(\bar{p})\bar{k}.$$

L satisfies P5(I) by 14.11 and 16.11. Γ_L satisfies P6 by 12.52. Then L satisfies P6(I) and P7(I) by 16.12. The fundamental lemma, 12.0, now proves that L satisfies all the postulates.

16.71 We call a correspondence satisfying all the postulates PR.

16.72 *Proof of 16.02:* Suppose $\text{Re}(Z(p_0)k, k) = 0$ for some $p_0 \in \Gamma_+$. Because this function of p is PR (16.12) we have

$$J_k(p) = (Z(p)k, k) \equiv 0.$$

Hence $k \in \mathbf{K}_1 = \mathbf{K}_{L0}$ (12.12, 12.55). Hence $[0, k] \in L(p)$ for every $p \in \Gamma_L$. That is

$$Z(p)k = 0 \quad \text{for } p \in \Gamma_L.$$

16.73 *Corollary:* If $Z(p_0)k = 0$ for some $p_0 \in \Gamma_+$, then $Z(p)k \equiv 0$. For the hypothesis here implies that of 16.72. This is the analog of 15.12; the result of 16.02 is stronger.

16.8 An important consequence of 16.7 is the

Lemma:* If $Z(p)$ is PR and of rank m , then there exists a real coordinate frame in which the matrix $[Z(p)]$ is an $m \times m$ nonsingular PR matrix $[Z_1(p)]$ bordered by zeros.

* Proved by Cauer⁵.

Proof: Consider the PR correspondence L defined by $Z(p)$. Then $\mathbf{V}_{L0} = 0$, because $Z(p)0 = 0$ for every $p \in \Gamma_L$. Consider the real frame of 13.02. $[Z(p)]$ in this frame takes any of k_{r+m+1}, \dots, k_n into 0 because these span \mathbf{K}_{L0} . Within \mathbf{K}_2 , $[Z(p)]$ must describe the same operation as the $[Z_1(p)]$ of 12.54. Because $[Z(p)]$ is symmetric the lemma follows.

XVII. THE JUXTAPOSITION OF CORRESPONDENCES

17.0 This section and the next will consider ways of constructing new correspondences from old. This will provide the basis of the necessity proof of Section 19.

17.01 It is obvious that if two physical networks are set side by side and their accessible terminals regarded as the terminals of a single larger network, that enlarged network is again a physical network. This is the gist of the present section.

17.1 Suppose that

$$\mathbf{V} = \mathbf{V}_1 \oplus \mathbf{V}_2, \quad \mathbf{K} = \mathbf{K}_1 \oplus \mathbf{K}_2,$$

where $\mathbf{K}_i = \mathbf{V}_i^*$ and all spaces are real (10.6). Let E_1 project on \mathbf{V} along \mathbf{V}_2 (Halmos⁹, par. 33) and $E_2 = 1 - E_1$ project on \mathbf{V}_2 along \mathbf{V}_1 . Then E_i^* projects on \mathbf{K}_i along \mathbf{K}_j , $j \neq i$ (Halmos⁹, loc. cit.). It is easily verified that $E_i = \bar{E}_i$, $E_i^* = \bar{E}_i^*$, from the analog of 14.02 for dimensionless operators.

Considering \mathbf{V}_i and \mathbf{K}_i as separate spaces, let L_i be a geometrical linear correspondence between them with frequency domain Γ_i , $i = 1, 2$.

Consider the correspondence L between \mathbf{V} and \mathbf{K} defined by

(i) The frequency domain $\Gamma_L = \Gamma_1 \cap \Gamma_2$

(ii) $[v, k] \in L(p)$ if and only if $[E_i v, E_i^* k] \in L_i(p)$, $i = 1, 2$.

In (ii), of course, we regard $E_i v$ and $E_i^* k$ as elements of \mathbf{V}_i , \mathbf{K}_i .

17.11 L so defined is called the juxtaposition of L_1 and L_2 .

17.2 *Lemma:* L is PR if and only if each of L_1 and L_2 is PR.

17.21 *Proof of "if":* It is clear that L satisfies P1 and P2. Further notation is now simplified if we put $L_1 = M$, $L_2 = N$. Consider the manifolds

$$\mathbf{V}_M \oplus \mathbf{V}_N, \quad \mathbf{V}_{M0} \oplus \mathbf{V}_{N0}, \quad \mathbf{K}_M \oplus \mathbf{K}_N, \quad \mathbf{K}_{M0} \oplus \mathbf{K}_{N0},$$

where $\mathbf{V}_M \subseteq \mathbf{V}_1$ is the manifold of voltages admitted by $L_1 = M$ considered as a correspondence between \mathbf{V}_1 and \mathbf{K}_1 , and \mathbf{V}_{M0} the manifold

of voltages $v \in \mathbf{V}_1$ such that $[v, 0] \in L_1(p)$ for all $p \in \Gamma_1$. Dual definitions for \mathbf{K}_M , \mathbf{K}_{M0} , and symmetrical ones for \mathbf{V}_N , \dots , \mathbf{K}_{N0} need not be repeated.

It is clear from these definitions that the four manifolds above are, in the order listed, the manifolds

$$\mathbf{V}_L, \mathbf{V}_{L0}, \mathbf{K}_L, \mathbf{K}_{L0}$$

for L . Now, for example,

$$(\mathbf{K}_{L0})^0 = (\mathbf{K}_{M0} \oplus \mathbf{K}_{N0})^0 = (\mathbf{K}_{M0})^0 \cap (\mathbf{K}_{N0})^0$$

by 10.6. This last manifold, in \mathbf{V} , is $(\mathbf{V}_M \oplus \mathbf{V}_2) \cap (\mathbf{V}_N \oplus \mathbf{V}_1)$, by P3 for M and N , and by 10.6. But by direct calculation

$$(\mathbf{V}_M \oplus \mathbf{V}_2) \cap (\mathbf{V}_N \oplus \mathbf{V}_1) = \mathbf{V}_M \oplus \mathbf{V}_N = \mathbf{V}_L.$$

The dual of this result then completes P3 for L .

P4 for L is immediate because the E_i and E_i^* are real.

The duality of the decompositions of \mathbf{V} and \mathbf{K} implies the identity

$$(v, k) = (E_1 v, E_1^* k) + (E_2 v, E_2^* k)$$

(that is $E_1 E_2 = E_2 E_1 = 0$, and dually. This is Halmos⁹, par. 33). All of P5, P6, and P7 for L follow at once from this identity.

17.22 The “only if” of 17.2 is a special case of the result of Section 18. Its proof will be deferred to 18.4.

17.23 It is obvious that the notion of juxtaposition and the lemma of 17.2 extend to juxtapositions of more than two correspondences.

17.3 Even without the “only if” part of 17.2, we have enough for the following characterization of PR correspondences:

Theorem: A correspondence L is PR if and only if it is the juxtaposition of

- (i) a correspondence defined by a nonsingular PR matrix between a \mathbf{V}_1 and a $\mathbf{K}_1 = \mathbf{V}_1^*$,
- (ii) a correspondence consisting of short circuits: that is of pairs $[0, k]$ for all $k \in \mathbf{K}_2$ and all p ,
- (iii) a correspondence consisting of open circuits: that is, of pairs $[v, 0]$ for all $v \in \mathbf{V}_3$ and all p .

Proof: If L is PR, the decomposition indicated is that of 13.1, 13.11, 13.12. If L is the juxtaposition indicated, then it is PR by 16.6 and the “if” in 17.1, provided the short and open circuits are PR correspondences. The verification of the postulates for these latter is easy and will be omitted.

17.31 The labor of considering PR correspondences instead of matrices has yielded the disappointingly simple result of 17.3. We have already been warned of this, however, by our knowledge of the properties of physical networks (2.9).

XVIII. THE OPERATION OF RESTRICTION

18.0 In addition to juxtaposition, which is an operation on correspondences clearly motivated by physical considerations, there is an operation, here called restriction, which has important use in the next section. There the physical meaning of the operation will become clear.

18.1 Let \mathbf{V} and $\mathbf{K} = \mathbf{V}^*$ be a pair of dual spaces. Let \mathbf{U} and $\mathbf{J} = \mathbf{U}^*$ be another pair. Suppose that C is a given fixed linear operation from \mathbf{J} to \mathbf{K} : given any $j \in \mathbf{J}$, there is a unique $k(j) \in \mathbf{K}$, written

$$k(j) = Cj,$$

such that if $k_r = Cj_r$, $r = 1, 2$, then

$$a_1 k_1 + a_2 k_2 = C(a_1 j_1 + a_2 j_2)$$

for any complex scalars a_1, a_2 .

18.11 Let $(v, k)_1$ denote the scalar product between \mathbf{V} and \mathbf{K} , and $(u, j)_2$ that between \mathbf{U} and \mathbf{J} . Given C , and any $v \in \mathbf{V}$, let us find that unique vector $u(v) \in \mathbf{U}$ for which

$$(u(v), j)_2 = (v, Cj)_1 \quad (1)$$

for every $j \in \mathbf{J}$. That such a vector $u(v)$ exists and is unique follows from 10.13 when we notice that the right-hand side of (1) defines a function conjugate linear in j . Now for fixed j , the right-hand side of (1) is linear in v , hence so also is the left side. That is, there is a linear operation C^* from \mathbf{V} to \mathbf{U} such that

$$u(v) = C^*v.$$

The following chart illustrates the situation:

$$\begin{array}{ccc} \mathbf{V} & & \mathbf{K} \\ C^* \downarrow & & \uparrow C \\ \mathbf{U} & & \mathbf{J} \end{array}$$

18.12 We suppose now that C takes real j into real k , i.e., that C is real. Then by (1)

$$\overline{(C^*v, j)_2} = \overline{(C^*v, j)_2} = \overline{(v, Cj)_1} = (\bar{v}, Cj)_1.$$

By comparison with (1), we have

$$\overline{C^*v} = C^*\bar{v}.$$

Hence C^* also takes real vectors into real vectors and is real.

18.2 Now let L be a PR correspondence between \mathbf{V} and \mathbf{K} . We define one, say M , between \mathbf{U} and \mathbf{J} , as follows: For each $p \in \Gamma_L$, let $M(p)$ consist of all pairs

$$[u, j]$$

such that $u = C^*v$ and

$$[v, Cj] \in L(p).$$

This definition can be illustrated by enlarging the chart of 18.11:

$$\begin{array}{ccc} \mathbf{V} & \xleftarrow{L} & \mathbf{K} \\ C^* \downarrow & & \uparrow C \\ \mathbf{U} & \xleftarrow{M} & \mathbf{J} \end{array}$$

The u 's corresponding to $j \in \mathbf{J}$ can be constructed by going around through C , L and C^* . This then defines a direct mapping from \mathbf{J} to \mathbf{U} .

18.21 We call the M defined by 18.2 a restriction of L , since its pairs are images under C^* and C^{-1} (which is not defined over all of \mathbf{K}) of a restricted set of pairs drawn from L .

18.22 Clearly there is a dual operation defined by an operator D from \mathbf{U} to \mathbf{V} . We might distinguish the operation of 18.2 by calling it a current restriction, its dual by calling it a voltage restriction.

18.23 The restriction M of L is defined by lists $M(p)$ which exist for any $p \in \Gamma_L$. The frequency domain of M has not yet been specified, however.

18.3 *Theorem:* If L is PR, then there is a frequency domain Γ_M for M such that M is PR.

Proof: P1 and P2 for M are evident at once, for any $p \in \Gamma_L$. The remainder of the proof is divided among 18.31, \dots , 18.37 below.

18.31 For P3, let \mathbf{J}_M be all $j \in \mathbf{J}$ such that $Cj \in \mathbf{K}_L$. Then, given $j \in \mathbf{J}_M$, for each $p \in \Gamma_L$ there is a v such that

$$[v, Cj] \in L(p),$$

whence

$$[C^*v, j] \in M(p).$$

Therefore $\mathbf{J}_M(p)$, the space of currents admitted by M at frequency p , coincides with the fixed \mathbf{J}_M at each $p \in \Gamma_L$.

Clearly \mathbf{J}_M is a real linear manifold.

18.32 Consider now $\mathbf{U}_{M0}(p)$: if $[u, 0] \in M(p)$, then there is a v such that $u = C^*v$ and

$$[v, C0] = [v, 0] \in L(p).$$

Hence $v \in \mathbf{V}_{L0}(p) = \mathbf{V}_{L0}$ for each $p \in \Gamma_L$. Therefore, for each $p \in \Gamma_L$,

$$\mathbf{U}_{M0}(p) \subseteq C^* \mathbf{V}_{L0}. \quad (2)$$

Now suppose, conversely, that $p \in \Gamma_L$ and $v \in \mathbf{V}_{L0} = \mathbf{V}_{L0}(p)$. Then $[v, 0] \in L(p)$. Now $0 = C0$, so $[v, C0] \in L(p)$. Hence $[C^*v, 0] \in M(p)$, so $C^*v \in \mathbf{U}_{M0}(p)$. This proves the inequality opposite to that of (2), so for $p \in \Gamma_L$

$$\mathbf{U}_{M0}(p) = C^* \mathbf{V}_{L0} = \mathbf{U}_{M0}, \quad (3)$$

a fixed space.

18.33 Now consider $(\mathbf{U}_{M0})^0$. If $j \in (\mathbf{U}_{M0})^0$, then

$$(u, j)_2 = 0$$

for every $u \in \mathbf{U}_{M0}$. That is, by (3),

$$(C^*v, j)_2 = (v, Cj)_1 = 0$$

for every $v \in \mathbf{V}_{L0}$. Therefore $Cj \in (\mathbf{V}_{L0})^0 = \mathbf{K}_L$, and $j \in \mathbf{J}_M$ by 18.31. That is, we have proved

$$\mathbf{J}_M \supseteq (\mathbf{U}_{M0})^0,$$

and, combining 18.31 with this and (3),

$$\mathbf{J}_M(p) = \mathbf{J}_M \supseteq (\mathbf{U}_{M0}(p))^0 = (\mathbf{U}_{M0})^0. \quad (4)$$

This is the weak form P3'(I) of 12.0 for M . It is as far as we can go with P3 at the moment.

18.34 Consider P4. If for $p \in \Gamma_L$ we have

$$[u, j] \in M(p)$$

then $[v, Cj] \in L(p)$ and $u = C^*v$. But then $[\bar{v}, C\bar{j}] \in L(\bar{p})$ and $\bar{u} = C^*\bar{v}$, by 18.12. Then however

$$[\bar{u}, \bar{j}] \in M(\bar{p})$$

by definition of M . This is P4.

18.35 Consider P5(I): if

$$[u_r, j_r] \in M(p),$$

where j_r is real, $r = 1, 2$, then

$$(u_r, j_s)_1 = (C^*v_r, j_s)_1 = (v_r, Cj_s)_1, \quad (5)$$

where $[v_r, Cj_r] \in L(p)$. Since Cj_r is real

$$(v_1, Cj_2)_1 = (v_2, Cj_1)_1$$

by P5(I) for L . This with (5) for $r \neq s$ proves P5(I) for M .

18.36 Fix a $j \in \mathbf{J}_M$ and for each $p \in \Gamma_L$ a $u(p)$ such that

$$[u(p), j] \in M(p).$$

Then $u(p) = C^*v(p)$ and

$$[v(p), Cj] \in L(p),$$

for some $v(p)$. Then as in (5) above

$$(u(p), j)_2 = (v(p), Cj)_1.$$

P6(I) and P7(I) for L then imply that P6(I) and P7(I) hold for M , using Γ_L for Γ_M in P6.

18.37 We now have M satisfying the hypotheses of 12.0. Therefore there is a Γ_M such that M satisfies all the postulates. This is 18.3.

18.4 *Proof of "only if" in 17.2:* Suppose that L between \mathbf{V} and \mathbf{K} is the juxtaposition of L_1 between \mathbf{V}_1 and \mathbf{K}_1 , L_2 between \mathbf{V}_2 and \mathbf{K}_2 . Let, say, $\mathbf{U} = \mathbf{V}_1$ and $\mathbf{J} = \mathbf{K}_1$. Let C be the identity map from \mathbf{K}_1 to \mathbf{K} : if $j \in \mathbf{J} = \mathbf{K}_1$, then Cj is just j considered as a vector in \mathbf{K} . Then C is real. It is easily computed that C^* is E_1 .

Consider the restriction M of L based on this C . Its pairs for $p \in \Gamma_M \subseteq \Gamma_L$ are all the pairs $[u, j]$ such that $j = E^*j \in \mathbf{K}_L$ and $u = Ev$, where

$$[v, j] \in L(p). \quad (6)$$

But then

$$[u, j] = [Ev, E^*j]$$

and this is in $L_1(p)$ by (6) and the definition of juxtaposition. Therefore the list $M(p)$ is contained in $L_1(p)$.

Suppose that $[u, j] \in L_1(p)$. We have $[0, 0] \in L_2(p)$ so by P2 and the defi-

nition of juxtaposition

$$[u, j] \epsilon L(p).$$

But then $j = E^*j$, $u = Eu$, and by definition of M

$$[u, j] \epsilon M(p).$$

Therefore for every $p \in \Gamma_M$, $M(p) = L_1(p)$. Therefore there is a frequency domain (Γ_M) for L_1 such that L_1 is PR.

XIX THE NECESSITY PROOF

19.0 Fortunately for this section, those parts of network theory which we require have recently been very succinctly stated by J. L. Synge¹². We shall paraphrase them here, referring the reader to the source¹² for details of definition.

19.01 First, we observe that in Cauer's definition⁵, which we shall repeat in detail below, an ideal transformer with m windings is a $2m$ -pole whose terminal pairs are the termini of the respective windings.

A system of m coupled coils is a $2m$ -pole with similarly defined terminal pairs.

19.02 Given a $2n$ -pole \mathbf{N} which is a finite passive network, let us adjoin ideal transformers as in Figure 1(b). We then draw the ideal graph of this network. Adjoin to the graph ideal *generator branches* $\gamma_1, \dots, \gamma_n, \gamma_r$ between T_r and T'_r , $1 \leq r \leq n$. Let β_r be the ideal branch representing the transformer winding between T_r and T'_r , $1 \leq r \leq n$. Enumerate the remaining branches of the graph $\beta_{n+1}, \dots, \beta_b$.

19.03 The branch γ_r is in a mesh with β_r and no other branches. Let us call this the r -th external mesh. Any basic set of meshes must include each of these.

19.04 Let ℓ_1, \dots, ℓ_n be the currents in the generator branches, k_1, \dots, k_b the currents in the branches β_1, \dots, β_b and

$$[\ell] = [\ell_1, \dots, \ell_n, k_1, \dots, k_b], \quad [k] = [k_1, \dots, k_b].$$

Let w_1, \dots, w_n be the voltages across the generator branches, v_1, \dots, v_b the currents in the β_1, \dots, β_b and

$$[w] = [w_1, \dots, w_n, v_1, \dots, v_b], \quad [v] = [v_1, \dots, v_b].$$

19.05 Let us choose a basic set of meshes, let j_1, \dots, j_s be the respective mesh currents, and

$$[j] = [j_1, \dots, j_s].$$

Let

$$[u] = [u_1, \dots, u_s]$$

be the s -tuple of mesh voltages. We suppose that $j_1, \dots, j_n, u_1, \dots, u_n$ refer respectively to the n external meshes. (Cf. 19.03.)

19.06 The results of Synge¹² can now be stated as follows:

There exists a real constant matrix $[C_1]$ of s columns and $b + n$ rows (having, in fact, elements which are $+1$, -1 , or 0) such that for any $[j]$

$$[\ell] = [C_1][j] \quad (1)$$

is a set of branch currents satisfying Kirchoff's node law, and for any $[w]$

$$[u] = [C_1]'[w] \quad (2)$$

is a set of mesh voltages satisfying Kirchoff's mesh law. Furthermore, given any $[\ell]$ which satisfies the node law, there is a $[j]$ such that (1) holds.

19.07 If we interpret the $[\ell]$, $[j]$, etc., as representations in real bases then $[C_1]$ is real and $[C_1]' = [C_1]^*$.

19.08 The matrix $[C_1]$ has the form

$$[C_1] = \begin{array}{|c|c|} \hline C_2 & 0 \\ \hline 0 & C \\ \hline \end{array}$$

where $[C_2]$ is an $n \times n$ diagonal matrix (having diagonal elements ± 1 , in fact).

Proof: By construction, j_1, \dots, j_n are mesh currents in the external meshes. These are then equal, save for sign, to the currents ℓ_1, \dots, ℓ_n in the generator branches.

19.09 By 19.08, (1), and the definitions in 19.04,

$$[k] = [C][j], \quad [u] = [C]'[v],$$

and by 19.07, $[C]' = [C]^*$.

19.1 Let us suppose that we have enumerated the branches $\beta_{n+1}, \dots, \beta_b$ in 19.02 in such a way that $\beta_{n+1}, \dots, \beta_c$ are all the two poles in the graph, $\beta_{c+1}, \dots, \beta_d$ are all the branches containing coils which are magnetically coupled, and $\beta_{d+1}, \dots, \beta_b$ the remaining ideal branches of ideal transformers.

Let $[Z_d(p)]$ be the $(d - n) \times (d - n)$ impedance matrix relating the voltages across the branches $\beta_{n+1}, \dots, \beta_d$ to the currents in them when

we consider the individual two-poles and the system of coupled coils as separate unconnected networks. Then $[Z_d(p)]$ is composed of a $(c - n) \times (c - n)$ diagonal matrix in the upper left field and a $(d - c) \times (d - c)$ matrix in the lower right, with zeros elsewhere.

19.11 The diagonal part of $[Z_d(p)]$ has elements drawn from the following list:

- (i) $f(p) = \rho$
- (ii) $f(p) = \delta p$
- (iii) $f(p) = \lambda p$

where ρ, δ, λ are non-negative constants, possibly different for each branch.

19.12 It is shown in texts on electromagnetic theory that the matrix representing a system of coupled coils is of the form

$$p[G],$$

where $[G]$ is a real, constant, symmetric, and semi-definite matrix. The lower right field of $[Z_d(p)]$ is then such a matrix.

19.13 It is obvious from this description that $[Z_d(p)]$ is PR. It therefore describes a PR correspondence between $(d - n)$ -tuples of current and voltage.

19.2 We must at last consider ideal transformers in detail. Let \mathbf{V}_1 and \mathbf{K}_1 be m -dimensional spaces represented as aggregates of m -tuples.

Let $\rho_1, \rho_2, \dots, \rho_m$ be m real numbers. Let \mathbf{V}_T consist of all m -tuples $[a] = [a_1, \dots, a_m]\epsilon\mathbf{V}_1$ such that

$$\frac{a_1}{\rho_1} = \frac{a_2}{\rho_2} = \dots = \frac{a_m}{\rho_m}.$$

We interpret these relations as follows:

- (a) If any $\rho_r = 0$, then $a_r = 0$
- (b) If any two ρ_r, ρ_s are not zero, then

$$\frac{a_r}{\rho_r} = \frac{a_s}{\rho_s}$$

- (c) If only one $\rho_r \neq 0$, then a_r is arbitrary.

Let \mathbf{K}_T consist of all m -tuples $[b] = [b_1, \dots, b_m]\epsilon\mathbf{K}_1$ such that

$$\rho_1 b_1 + \rho_2 b_2 + \dots + \rho_m b_m = 0.$$

\mathbf{V}_T and \mathbf{K}_T are linear manifolds.

Let $[L_T]$ be the concrete linear correspondence defined by the list $[L_T](p)$ which consists for each complex p of all pairs $[[a], [b]]$ where $[a] \in \mathbf{V}_T$, $[b] \in \mathbf{K}_T$.

The correspondence described by $[L_T]$ is what Cauer⁵ defines as an ideal transformer. He shows, loc. cit., how it can be defined as the limiting case of a physical transformer.

There is also a dual kind of device, described by a correspondence admitting all $[b] \in \mathbf{K}_1$ for which

$$\frac{b_1}{\lambda_1} = \frac{b_2}{\lambda_2} = \dots = \frac{b_m}{\lambda_m}$$

and all $[a] \in \mathbf{V}_1$ for which

$$\lambda_1 a_1 + \dots + \lambda_m a_m = 0.$$

This also is an ideal transformer obtainable as a limiting case of a physical one.

19.21 The correspondence L_T is PR.

Proof: We observe that $\mathbf{V}_T = (\mathbf{K}_T)^0$, for let $[a] \in \mathbf{V}_T$, $[b] \in \mathbf{K}_T$, and let t be the common value of the a_r/ρ_r . Then

$$(a, b) = \Sigma a_r \bar{b}_r = t \Sigma \rho_r \bar{b}_r = \overline{t(\Sigma \rho_r b_r)} = 0.$$

The postulates are now all easily proved. We omit the details.

19.3 Let \mathbf{V} and \mathbf{K} be b -dimensional spaces. We interpret the b -tuples $[v]$ and $[k]$ of 19.04 as representing vectors $v \in \mathbf{V}$, $k \in \mathbf{K}$ in a real frame.

Let L be the correspondence between \mathbf{V} and \mathbf{K} formed by juxtaposing

(i) the correspondence described by $[Z_d(p)]$ relating components with indices in the range $n + 1$ to d ,

(ii) the several correspondences described by ideal transformers, relating components with indices in the ranges 1 to n and $d + 1$ to b .

L is PR because it is the juxtaposition of PR correspondences.

19.31 Let \mathbf{U} and \mathbf{J} be $s - n$ -dimensional spaces. We interpret the $[u]$ and $[j]$ of 19.04 as representing $u \in \mathbf{U}$, $j \in \mathbf{J}$ in a real frame.

19.32 Let C be the operation from \mathbf{J} to \mathbf{K} whose matrix in our chosen frames is $[C]$. Then C^* operates from \mathbf{V} to \mathbf{U} with the matrix $[C]^* = [C]'$. By these definitions, C is real. Let M be the correspondence between \mathbf{U} and \mathbf{J} obtained by restricting L with C . Then there is a frequency domain Γ_M such that M is PR (18.3).

19.4 By 19.09, $[M]$ in our chosen frame is the correspondence established between mesh currents and mesh voltages by the network of the

$2n$ -pole \mathbf{N} . When this network operates as a $2n$ -pole, the only mesh voltages which are not zero are those relating to the external meshes, since there are no internal sources of voltage. We must now account for this.

19.41 Let $\mathbf{V}_2, \mathbf{K}_2$ be n -dimensional spaces. Choose a real frame and let D be the operation which takes

$$[a_1, \dots, a_n] \epsilon \mathbf{V}_2 \quad (3)$$

into

$$[a_1, \dots, a_n, 0, \dots, 0] \epsilon \mathbf{U} \quad (4)$$

in the frame of 19.31. Then D is real and D^* in the chosen frames takes

$$[b_1, \dots, b_s] \epsilon \mathbf{J} \quad (5)$$

into

$$[b_1, \dots, b_n] \epsilon \mathbf{K}_2. \quad (6)$$

19.42 We interpret the n -tuples (3) and (6) as voltages and currents in the external meshes of \mathbf{N} . Their relations to (4) and (5) are consistent with this interpretation.

Let us restrict M by D , to get a correspondence M_1 between \mathbf{V}_2 and \mathbf{K}_2 . In our chosen frame, the passage to $[M_1]$ corresponds, by (3) and (4) of 19.41, to considering mesh voltages in \mathbf{N} which vanish for every internal mesh, and, correspondingly letting the mesh currents adjust themselves to this situation. We of course observe only the external mesh currents (6).

19.43 M was PR. So, therefore is M_1 (18.3 dual). Since $[M_1]$ is the correspondence established by the physically realizable $2n$ -pole \mathbf{N} , the necessity of P1, \dots , P7 for formal realizability is established.

XX. APPENDIX TO PART I

20.0 We must prove 7.22 and those assertions of 10.6 which are not covered in Halmos⁹. These concern reality.

20.1 Let \mathbf{V}_1 be a real manifold and

$$\mathbf{V} = \mathbf{V}_1 \oplus \mathbf{V}_2, \quad \mathbf{K} = \mathbf{K}_1 \oplus \mathbf{K}_2$$

where $\mathbf{K}_1 = (\mathbf{V}_2)^0$, etc. The basis (14) of 10.6 exists by Halmos⁹, par. 19. We show that it can be chosen to be real. We have linearly independent vectors

$$v_1, \dots, v_r, v_{r+1}, \dots, v_n,$$

where the first r span \mathbf{V}_1 , the last $n - r$, \mathbf{V}_2 . Let

$$v_s = u_s + iw_s, \quad 1 \leq s \leq n,$$

where u_s, w_s are real (10.42). Since \mathbf{V}_1 is real and a linear manifold,

$$u_s = \frac{1}{2}(v_s + \bar{v}_s)\epsilon\mathbf{V}_1, \quad 1 \leq s \leq r,$$

and, similarly, $w_s\epsilon\mathbf{V}_1$, $1 \leq s \leq r$. Among the $2n$ real vectors

$$u_1, u_2, \dots, u_r, w_1, \dots, w_r, u_{r+1}, \dots, u_n, w_{r+1}, \dots, w_n, \quad (1)$$

the first $2r$ are in \mathbf{V}_1 , and they span \mathbf{V}_1 because the v_s , $1 \leq s \leq r$, can be constructed from them. The whole list (1) spans \mathbf{V} , because from it all the v_s , $1 \leq s \leq n$, can be constructed. Since the $v_s\epsilon\mathbf{V}_2$ do not use in their construction any of the first $2r$ vectors (1), it follows that the last $2(n - r)$ vectors in that list must contain a set spanning \mathbf{V}_2 . The reality of the vectors (1) then establishes the existence of a real basis, say,

$$v'_1, \dots, v'_r, v'_{r+1}, \dots, v'_n \quad (2)$$

which provides a basis in \mathbf{V}_1 and \mathbf{V}_2 .

20.11 We now have 7.22. The unique dual basis

$$k'_1, \dots, k'_n$$

to (2) is real by 10.41. Hence all of $\mathbf{V}_1, \mathbf{V}_2, \mathbf{K}_1, \mathbf{K}_2$ are real. The proof of 10.6 is then complete.

20.2 If in a real basis (2) (dropping primes)

$$v = a_1v_1 + a_2v_2 + \dots + a_nv_n,$$

that is, if

$$[v] = [a_1, \dots, a_n],$$

then by (5) of 10.3

$$\bar{v} = \bar{a}_1v_1 + \dots + \bar{a}_nv_n,$$

hence

$$[\bar{v}] = [\bar{a}_1, \dots, \bar{a}_n].$$

The geometrical conjugation of 10.3 is therefore simply the concrete one of 7.2 in any real basis. This proves the remark of 10.35.

BIBLIOGRAPHY

1. M. Bayard, "Synthèse des Réseaux Passifs a un Nombre Quelconque de Paires de Bornes Connaissant Leurs Matrices d'Impedance ou d'Admittance," *Bulletin, Société Française des Electriciens*, **9**, 6 series, Sept. 1949.
2. O. Brune, *Jour. Math. and Phys., M.I.T.*, **10**, Oct. 1931, pp. 191-235.
3. W. Cauer, *Ein Reaktanztheorem*, *Sitzungsberichte Preuss. Akad. Wissenschaft*, Heft 30, 32, 1931.
4. W. Cauer, "Die Verwirklichung von Wechselstromwiderständen vorgeschriebener Frequenzabhängigkeit," *Archiv für Elektrotechnik*, **17**, 1926.
5. W. Cauer, "Ideale Transformatoren und Lineare Transformationen," *Elektrische Nachrichten-Technik*, **9**, May, 1932.
6. S. Darlington, *Journal of Mathematics and Physics, M.I.T.*, **18**, No. 4, Sept. pp. 257-353.
7. R. M. Foster, *Bell System Tech. J.*, April, 1924, pp. 259-267.
8. C. M. Gewertz, *Network Synthesis*, Baltimore, 1933.
9. P. R. Halmos, *Finite Dimensional Vector Spaces*, Princeton, 1942.
10. Y. Oono, "Synthesis of a Finite $2n$ -Terminal Network by a Group of Networks Each of Which Contains Only One Ohmic Resistance," *Jour. Inst. Elec. Comm. Eng. of Japan*, March, 1946. Reprinted in English in the *Jour. Math. and Phys., M.I.T.*, **29**, Apr., 1950.
11. Y. Oono, "Synthesis of a Finite $2n$ -Terminal Network as the Extension of Brune's Theory of Two-Terminal Network Synthesis," *Jour. Inst. Elec. Comm. Eng. of Japan*, Aug., 1948.
12. J. L. Synge, "The Fundamental Theorem of Electrical Networks," *Quarterly of Applied Mathematics*, **9**, No. 2, July, 1951.
13. R. Bott, and R. J. Duffin, "Impedance Synthesis Without the Use of Transformers," *Jour. Appl. Phys.*, **20**, Aug., 1949, p. 816.
14. H. W. Bode, *Network Analysis and Feedback Amplifier Design*, New York, 1945.

An Application of Boolean Algebra to Switching Circuit Design

BY ROBERT E. STAEHLER

(Manuscript received January 10, 1952)

This paper discusses the application of switching (Boolean) algebra to the development of an all-relay dial pulse counting and translating circuit employing the minimum number of relays. An attempt is made to outline what appears to be the most promising method of obtaining beneficial results from the use of the algebra in the design of practical switching circuits.

INTRODUCTION

The demands made upon telephone switching systems in regard to improvements in handling capacity, speed, flexibility and economy are continually increasing. In order to meet design objectives enabling the fulfillment of these demands, switching circuits have of necessity become more and more complex and intricate. As certain types of relay switching circuits increase in complexity, the problem of control and output contact network design becomes more and more laborious and time consuming. This is especially true in those circuits in which an attempt has been made to achieve the ultimate in efficiency and economy in that the number of relays used therein approaches the absolute minimum necessary to provide the required number of distinct output combinations. In this type of near-minimum combinational or sequential relay circuit there are numerous parallel control and output contact paths which thread through the same relays repeatedly, thereby causing the individual relay contact loads to become relatively large. Thus the designer's problem becomes that of first developing a workable control and output contact network and then manipulating and minimizing contacts within that network so that the maximum number of contacts used on any one relay is within that permissible on any commercially available relay having the necessary speed characteristics.

Even in those combinational and sequential relay circuits which are not near-minimum and therefore probably have fairly light individual relay contact loads, there are, of course, advantages to be gained by using the least number of contacts possible. Although the initial cost per additional contact (assuming that a few added contacts per relay will not impair the relay speed or space characteristics to an extent that the circuit requirements are not met) is almost negligible, there are other

economic savings possible. Since each contact must be connected to the remainder of the contact network, minimizing contacts and consequently soldered connections means a saving in wiring time and labor. Furthermore, if the designer will manipulate the contacts so that the relays can be chosen from a comparatively few standardized codes, which are in large demand, it is possible to avoid the expensive stockpiling of numerous special designs having only a limited demand. In addition, using the least number of contacts minimizes the focal points of most relay circuit failures which are the contacts themselves (i.e., dirty or worn contacts).

It might also be noted at this point that electronic combinational or sequential circuits usually require electronic gating networks to perform functions which are completely analogous to those of relay contact networks. Hence, the same problem of minimization exists. However, in electronic circuits, gate minimization is even more advantageous since the cost per additional electronic gate is much higher than the cost per additional relay contact.

It is rather obvious that the multiplicity of paths in most combinational and sequential circuits can cause their design to become an extremely difficult and time consuming problem if the contact paths are developed with the aforementioned considerations in mind.

The circuit designer's usual approach to the solution of such contact minimization and manipulation problems is that of inspection. The method of inspection presupposes a background of considerable experience in that the designer must recognize certain contact network arrangements that may allow further rearrangements and thereby he must mentally develop his own rules. In order to check on any of his manipulations he must repeatedly redraw the network during this inspection design process. It is evident that this is often a long and tedious method and, depending on the skill of the designer, may or may not result in an optimum or even adequate solution.

Suitable contact network arrangements often appear only after consideration of several alternative schemes and the rearrangements of the network interconnections of these schemes. Realization of this makes it quite evident that any means of obtaining and comparing these various schemes quickly and with a mathematical accuracy which does not require continuous checking of network paths permits a more rapid and complete exploration of the particular problem. Switching algebra, first codified by C. E. Shannon¹, is the systematic application of G. Boole's²

¹ C. E. Shannon, *A Symbolic Analysis of Relay and Switching Circuits*, Trans. AIEE, **57**, 1938.

² G. Boole, *The Mathematical Analysis of Logic* (Cambridge 1847) and *An Investigation of the Laws of Thought* (London 1854).

"Algebra of Logic" to switching circuits and is just such a means. It is a tool which can be used to investigate the complex combinational and sequential networks to determine satisfactory contact arrangements or reject unsatisfactory ones with a minimum of time and effort. It should be emphasized, however, that as with any tool, satisfactory results depend upon the judgment, ingenuity and logical reasoning of the user. Furthermore, as will be evident from the following development, switching (Boolean) algebra in its present state is not to be considered entirely selfsufficient but, for the most beneficial results, should be applied, when warranted, in conjunction with inspection techniques so that the latter may fill in any limitations in the algebra techniques which have not been completely systematized as yet due to the newness of this field.

The problem of solving the contact requirements of a minimum relay dial pulse counting and translating circuit recently developed as a component of the originating register of the No. 5 Crossbar System will be used as a means of illustrating the practical use now being made of switching algebra and of indicating exactly where the application of the algebra enters the design problem.

BASIC DIAL PULSE COUNTER REQUIREMENTS

The primary function of the originating register is to receive pulse signals representing digits from a telephone dial or similar calling device and to store a record of the digits in a form suitable for use by an external circuit. The dial pulse counting and translating circuit, an integral part of the originating register, is oriented with respect to other parts of the register by the block diagram of Fig. 1. The *L* relay is the pulse detecting relay. When the subscriber's switchhook contact is closed due to the lifting of the phone, the originating register is connected to the line and the *L* relay is operated. Thereafter it follows the breaks and makes of the subscriber's dial and feeds these repeated dial pulses into the counter. After the pulses are counted they are translated to a new code. In switching systems it is advantageous to translate from the basic dial ten pulse decimal code to a "two out of five" self-checking code. In this latter code any single error within the circuit will result in either one or three relays operated in the associated storage circuit rather than two and thus an error can readily be detected. The output of the translator is fed via a steering circuit to the register or storage circuit. The slow release *RA* relay is the pulse train detecting relay which holds between the individual pulses of a digit and releases only at the end of the pulse train. When it releases it activates the translating circuit and thereby transfers the translated code information to the storage circuit. The *RA1* relay

in operating terminates the output from the translator and simultaneously releases the relays in the counter to prepare it for the next digit.

Specific requirements imposed by the originating register circuit necessitate the counting of one to eleven pulses; the use of a driving source consisting of a single break-make (or transfer) contact with ground on the armature spring; and outputs as follows:

1. Count of 1 through 10: ground on two of the 0, 1, 2, 4, 7 output leads in the combination corresponding to the count.
2. Count of 10: ground on the Z0 lead.
3. Count of 11: ground on the 0 lead only (this is a trouble-detecting feature).

In addition, the design of the steering and register-storage circuit requires that no output leads be connected together until the second pulse is received. Furthermore, each relay is limited to a combination of simple make and break contacts not exceeding a total of twelve. This utilizes the maximum number of springs obtainable on presently available relays and also avoids the larger armature gaps imposed by transfers which would result in a reduction in the relay speed of operation. Speed requirements also do not permit the use of shunt release in the circuit operation.

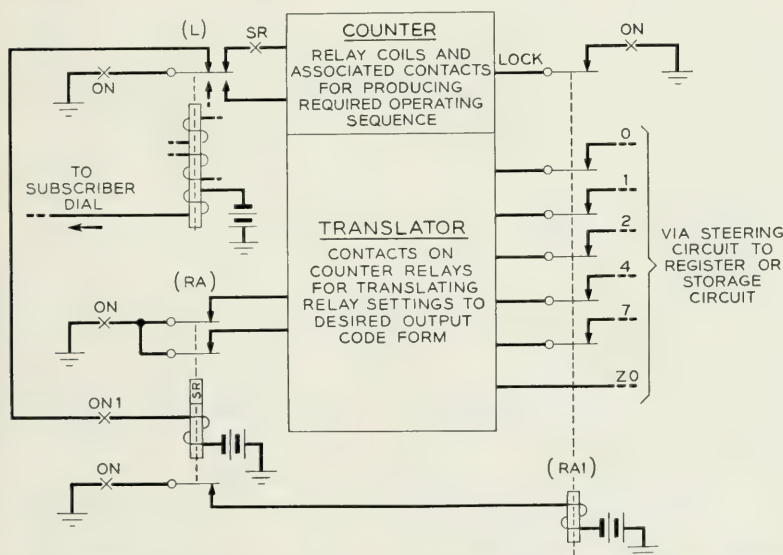


Fig. 1—The schematic of a portion of a dial pulse register circuit for counting decimal code pulses and translating them to "two out of five" signals. (In the symbolism used in the illustrations a cross indicates a "make" contact and a vertical bar indicates a "break" contact.)

THEORY OF A MINIMUM RELAY COUNTER

The counting circuit under consideration does not contemplate the use of any circuit elements other than relays that react to the beginning or end of a pulse. Therefore it must establish a distinct combination of relays operated or released during and between successive pulses. The minimum number of ordinary "two-position" relays, R , required to count P pulses can be obtained from the expressions (1) $2P \leq 2^R$ if the counter is to lock up during, or recycle after, the last pulse or (2) $2P \leq 2^R - 1$ if the counter is to lock up after the last pulse.

The usual counting circuit used for determining the number of pulses in a dial train is required to count ten pulses, however there are certain advantages in regard to trouble indications if the counter counts eleven pulses. In either case the minimum number of relays necessary, according to the preceding formulae, is five. It should be noted that the ease with which this minimum number can be attained depends upon whether the input is derived from a single, double or transfer contact source.

DETERMINATION OF OPERATING SEQUENCE

Having determined that the minimum number of relays necessary is five, the first step in design is to develop an operating sequence pattern from the resulting 2^5 or 32 possible relay combinations. These combinations may be utilized in any order deemed desirable to obtain the 23 distinct combinations needed to differentiate between eleven pulses (22 for the eleven makes and breaks plus an all-relays-normal combination). In this phase of the design switching algebra is not involved. The optimum sequence to meet a particular set of requirements can only be determined by repeated trials guided by an intimate knowledge of objectives.

Initial studies, made by Joseph Michal, of various possible sequence patterns for a five relay circuit, including those having a three relay "ring" followed by two auxiliary relays and those having a two relay pulse divider followed by three auxiliary relays, resulted in the conclusion that the latter approach was the most fruitful. The sequence pattern adopted is shown in detail in Table I. The pattern is extended through 12 pulses, and it can be seen that the nature of the sequence is such that this employs all 32 combinations of the 5 relays. Several of these are transient and occur during part of a pulse or inter-pulse interval. Examination of the tail end of the sequence indicates that it will be simpler to design on the basis of a full 12 pulses than attempt to block at the end of the 11 pulses specified by the requirements. If trouble con-

TABLE I
 SEQUENCE OF OPERATION

	Pulsing Relay <i>L</i>	Counting Relays					Relay Combination	Two out of Five Code
		<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>		
Seizure	0	1	1	1	1	1	1	
1st pulse	1	0	1	1	1	1	2	
	0	0	0	1	1	1	3	0,1
2nd pulse	1	1	0	1	1	1	4	
	1	1	0	0	1	1	5	
	0	1	1	0	1	1	6	0,2
3rd pulse	1	0	1	0	1	1	7	
	0	0	0	0	1	1	8	
	0	0	0	0	0	1	9	1,2
4th pulse	1	1	0	0	0	1	10	
	0	1	1	0	0	1	11	0,4
5th pulse	1	0	1	0	0	1	12	
	1	0	1	1	0	1	13	
	0	0	0	1	0	1	14	1,4
6th pulse	1	1	0	1	0	1	15	
	0	1	1	1	0	1	16	
	0	1	1	1	0	0	17	2,4
7th pulse	1	0	1	1	0	0	18	
	0	0	0	1	0	0	19	0,7
8th pulse	1	1	0	1	0	0	20	
	1	1	0	0	0	0	21	
	0	1	1	0	0	0	22	1,7
9th pulse	1	0	1	0	0	0	23	
	0	0	0	0	0	0	24	
	0	0	0	0	1	0	25	2,7
10th pulse	1	1	0	0	1	0	26	
	0	1	1	0	1	0	27	4, 7-ZO
11th pulse	1	0	1	0	1	0	28	
	1	0	1	1	1	0	29	
	0	0	0	1	1	0	30	0
12th pulse	1	1	0	1	1	0	31	
	0	1	1	1	1	0	32	0

Total of $2^5 = 32$ combinations used.

Note: 0 is used to indicate that the relay listed at the head of the column is operated, and 1 is used to indicate that the relay is released.

ditions introduce pulses beyond 12, the circuit will without difficulty recycle through combinations corresponding to pulses 11 and 12.

Table I also indicates the leads which must be grounded in order to provide the translations to the "two out of five" and "single lead" codes.

The characteristics of this circuit may be summarized as follows: It contains only five relays which is the absolute minimum necessary. It

uses all 32 of its available combinations. Its control and translating job is complex enough to indicate the need for a considerable number of contacts and hence the need for extensive contact manipulation to minimize and distribute these contacts.

It is apparent that a great deal of time would be necessary to accomplish this manipulation by inspection methods, thereby indicating the need for an additional tool such as switching algebra to assist the designers in this task.

ALGEBRAIC METHODS APPLIED TO CONTROL CIRCUIT

The sequence of operations of Table I is used as the starting point in the application of the algebra. The exact calculations necessary to develop the control and translating circuit by this means are shown in detail later. However, the individual steps in the solution might well be outlined here. First, the design of the control and translating networks will be regarded as separate problems. In theory these can be integrated together, but the resultant network is likely to be so complex that understanding and maintenance of the circuit would suffer. Each of the two networks can be individually considered as a multi-terminal network of the single input type. That is, the control network is an associated set of contacts which connects a single ground input to the windings of five relays, and the translating network is an associated set of contacts which connects a single ground input to the six output leads. Since switching algebra is directly applicable to two-terminal networks rather than multi-terminal networks, the approach to this particular problem is of necessity somewhat indirect.

The most satisfactory method of attack is to develop first a two-terminal network for each of the output paths of the multi-terminal network under consideration. The two-terminal networks can be expressed algebraically and manipulated into their simplest form by means of the switching algebra theorems to be given later. The individual networks can then be inspected carefully, either in algebraic or circuit form, with the objective of combining them in the most advantageous fashion. It will be found, in general, that the simplest network configurations do not readily combine and that further manipulation is necessary to obtain an economical circuit. It is at this point that the algebra achieves its greatest utility, since its application permits the simple and rapid changing of a given two-terminal network into a large variety of different forms with mathematical assurance that circuit equivalence is maintained. Inspection of the networks in the several forms provides clues

as to the preferable combining forms and often indicates additional manipulations that might be desirable.

This network development is a combination of mathematics and integration by inspection. It is characterized by repeated trials of alternative forms and at no stage is there any definite assurance that the optimum circuit has been attained. However, the ease of manipulation provided by the algebra greatly enhances the probability of designing a better circuit than would be possible by inspection alone. In combining the two-terminal networks, care must be taken not to introduce "sneak" paths which improperly connect outputs together. The algebra usually offers means of introducing one or two additional contacts which permit combining networks and yet eliminate the adverse effects of the sneak paths.

The above procedure will now be carried out in detail with the switching algebra theorems that are used in all the following algebraic manipulations noted at the margin by the number which corresponds to the number of the theorem in the complete listing in Table II. This table is

TABLE II
SWITCHING (BOOLEAN) ALGEBRA*

Definitions	Postulates
Addition (+) = AND = Series	(1) $X = 0$ or $X = 1$, where X is a contact or a network.
Multiplication (\cdot) = OR = Parallel	(2a) $0 \cdot 0 = 0$ (2b) $1 + 1 = 1$ (3a) $1 \cdot 1 = 1$ (3b) $0 + 0 = 0$ (4a) $1 \cdot 0 = 0 \cdot 1 = 0$ (4b) $0 + 1 = 1 + 0 = 1$
Circuit States 0 = Closed Circuit 1 = Open Circuit	
Theorems	
(1a) $X + Y = Y + X$ (1b) $XY = YX$ (2a) $X + Y + Z = (X + Y) + Z$ $= X + (Y + Z)$ (2b) $XYZ = (XY)Z = X(YZ)$ (3a) $XY + XZ = X(Y + Z)$ (3b) $(X + Y)(X + Z) = X + YZ$ (4a) $X + X = X$ (4b) $XX = X$ (5a) $X + XY = X$ (5b) $X(X + Y) = X$ (6a) $(X)' = X'$ (6b) $(X')' = X$ (7a) $(X + Y + Z + \dots)'$ $= X' \cdot Y' \cdot Z' \cdot \dots$ (7b) $(X \cdot Y \cdot Z \cdot \dots)'$ $= X' + Y' + Z' + \dots$	(8a) $X' + X = 1$ (8b) $X'X = 0$ (9a) $0 + X = X$ (9b) $1 \cdot X = X$ (10a) $1 + X = 1$ (10b) $0 \cdot X = 0$ (11a) $(X + Y')Y = XY$ (11b) $XY' + Y = X + Y$ (12a) $(X + Y)(X' + Z)(Y' + Z)$ $= (X + Y)(X' + Z)$ (12b) $XZ + X'Y + YZ = XZ + X'Y$ (13) $(X + Y)(X' + Z) = XZ + X'Y$ (14a) $f(X) = A \cdot f(X)_{A=1, A'=0}$ $+ A' \cdot f(X)_{A=0, A'=1}$ (14b) $f(X) = [A + f(X)_{A=0, A'=1}]$ $[A' + f(X)_{A=1, A'=0}]$

* Reprinted from *The Design of Switching Circuits* by Keister, Ritchie and Washburn with the permission of D. Van Nostrand Co., Inc.

taken from *The Design of Switching Circuits* by Keister, Ritchie and Washburn*. The development of the algebraic expressions from the sequence of operations table will be in exact parallel to the methods suggested in the aforementioned text.

The symbolism adopted in the following development is basically that of using the notation A for all the make contacts on the A relay, and A' for all the break contacts on the A relay. Contacts or groups of contacts in series are related by the symbol of addition (+) and contacts or groups of contacts in parallel are related by the symbol for multiplication (\cdot) which may or may not be explicitly written, as in ordinary algebra. Therefore $(A + B')$ symbolizes a series contact path that is closed when the A relay is operated *and* the B relay is released, while (AB') symbolizes the parallel contact path that is closed when either A is operated *or* B is released. Switching algebra includes only two numerical values, 0 and 1, with the quantity 0 assigned to represent a closed path and 1 to represent an open path. For the tabular notation of Table I, 0 is used to indicate that the relay listed at the head of the column is operated and 1 is used to indicate that the relay is released.

As stated earlier, the present application of switching algebra utilizes the sequence of operation chart of Table I. The operate and release combinations for controlling the A , B , C , D and E relays can be selected from this table by observing where each relay to be controlled changes state. For example, the operate combination for relay D is relay combination 8 and the release combination for relay D is relay combination 24. It is not necessary to include the contacts of a relay in its own operate and release combinations. Note that the A and B relays which serve as a pulse divider can be controlled solely by the L relay and contacts on A and B without reference to C , D , E . However the C , D and E relays are internally controlled by all five counting relays. The development of all these control paths uses the following abbreviations:

$g(X)$ = operating combinations for the X relay

$r(X)$ = releasing combinations for the X relay

$h(X)$ = holding combinations for the X relay

X = make contact on the X relay

Furthermore as expressed by theorem (6a and 6b) the negative of a contact network X is defined as a network which is a closed path under all conditions for which X is open, and is open under those conditions

* D. Van Nostrand, 1951. The Bell Telephone Laboratories Series.

for which X is closed. Hence $h(X)$ may be obtained from $r(X)$ by noting that $h(X)$ is the negative of $r(X)$. Therefore the entire control path of any relay can be expressed generally as

$$f(X) = g(X)[X + h(X)] = g(X)[X + (r(X))']$$

Thus for the A relay

$$\begin{aligned} g(A) &= L' + B' \\ r(A) &= L' + B \\ h(A) &= [L' + B]' = LB' \end{aligned} \tag{7a}$$

and

$$\begin{aligned} f(A) &= (L' + B')(A + LB') \\ &= (L' + B')(A + L)(A + B') \end{aligned} \tag{3b}$$

$$= (L + A)(L' + B') \tag{12a}$$

Also for the B relay

$$\begin{aligned} g(B) &= L + A \\ r(B) &= L + A' \\ h(B) &= [L + A']' = L'A \end{aligned} \tag{7a}$$

and

$$\begin{aligned} f(B) &= (L + A)(B + L'A) \\ &= (L + A)(B + L')(B + A) \end{aligned} \tag{3b}$$

$$= (L + A)(L' + B) \tag{12a}$$

The schematic forms of the A and B control circuits as represented by the above algebraic expressions are shown in Fig. 2a and 2b. Since the general requirements of the basic problem specify only one transfer on the L relay, only simple makes and breaks on the A and B relays and no shunt release paths (to avoid reduction in speed of operation), the combination of the above specific circuits is not possible without recourse to double windings. Another factor which affects the practical form of the circuit is the finite transit time of the L relay armature spring. Switching algebra presupposes instantaneous action of relay contacts and in certain cases, when the use of a break-make transfer is required, additional contacts are necessary to cover the open contact interval. The final circuit form, conceived by F. K. Low, is shown on Fig. 2c and uses an added A

contact. Algebraic equivalence of this circuit with the original is shown below.

$$\begin{aligned} f(A) &= (L'A + B')(L + A) \quad (\text{Fig. 2c}) \\ &= (L' + B')(A + B')(L + A) \end{aligned} \quad (3b)$$

$$= (L' + B')(L + A) \quad (12a)$$

$$\begin{aligned} f(B) &= (L'A + B)(L + A) \quad (\text{Fig. 2c}) \\ &= (L' + B)(A + B)(L + A) \end{aligned} \quad (3b)$$

$$= (L' + B)(L + A) \quad (12a)$$

For the C relay which operates and releases twice in the entire 32 combination cycle

$$\begin{aligned} g(C) &= (A' + B + D' + E')(A' + B + D + E) \\ r(C) &= [(A + B' + D + E')(A + B' + D' + E)] \\ h(C) &= [(A + B' + D + E')(A + B' + D' + E)]' \\ &= (A'BD'E + A'BDE') \end{aligned} \quad (7a, 7b)$$

and

$$\begin{aligned} f(C) &= [(A' + B + D' + E')(A' + B + D + E)] \\ &\quad [C + A'BD'E + A'BDE'] \\ &= [A' + B + (D' + E')(D + E)] \\ &\quad [C + A'B(D' + E')(D + E)] \quad (3a, 3b, 13) \\ &= [A' + B + (D' + E')(D + E)][C + A'B] \\ &\quad [C + (D' + E')(D + E)] \quad (3b) \\ &= [(A' + B)C + (D' + E')(D + E)][C + A'B] \quad (3b) \end{aligned}$$

The schematic circuit which the above represents is shown in Fig. 3a. Circuits of this type which use certain contacts more than once can some-

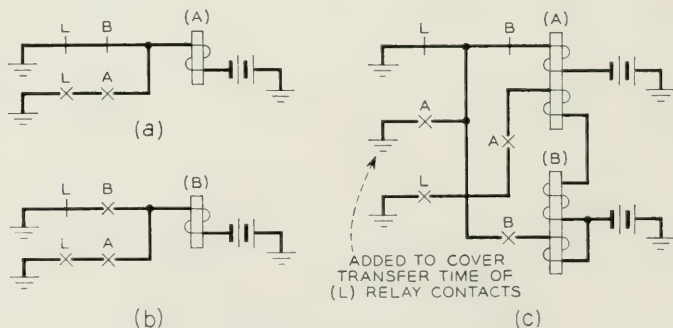


Fig. 2—Pulse divider of counting circuit.

times be drawn in bridge form with a consequent saving of contacts. One method is to manipulate the expression into a form which is known to be the series-parallel equivalent of a bridge. However, following usual algebraic procedures it is often difficult to recognize where this is possible. In the present case a method developed by G. R. Frost (not yet published) was used effectively. This resulted in the bridge circuit of Fig. 3b which has the series-parallel equivalent:

$$f(C) = [C + (D' + E')(D + E)][A' + B + (D' + E')(D + E)] \\ [C + A'][C + B]$$

By use of theorem (3b) this is seen to be equivalent to the previous expression for $f(C)$.

For the D relay which only operates and releases once in the entire cycle

$$g(D) = (A + B + C + E')$$

$$r(D) = (A + B + C + E)$$

$$h(D) = (A + B + C + E)'$$

$$= A'B'C'E'$$

and $f(D) = (A + B + C + E')(D + A'B'C'E')$.

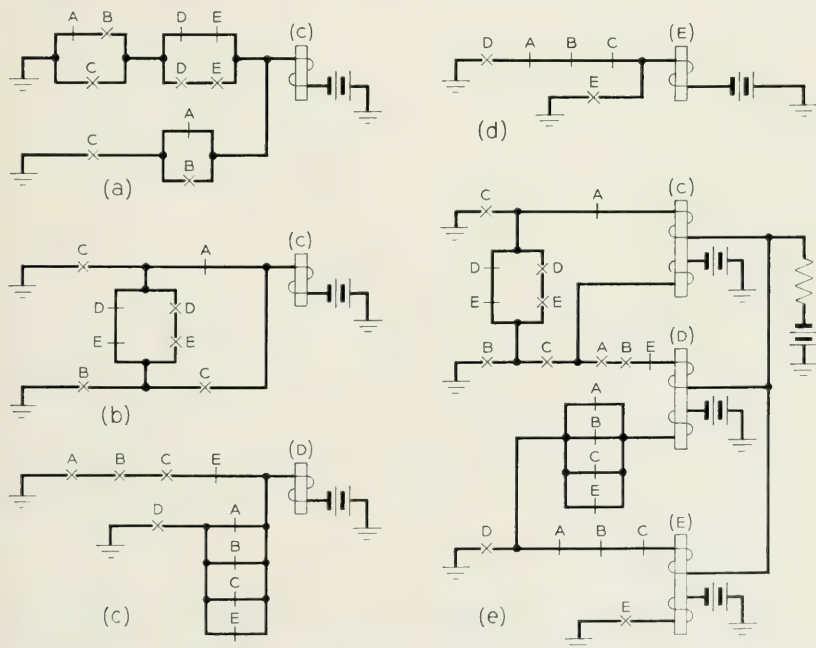


Fig. 3—Internal control of counting circuit.

By noting that $A + B + C$ is the negative of $A'B'C'$, this can be reduced to

$$f(D) = (A + B + C + E')(D + A'B'C') \quad (3b, 12a)$$

However, this transformation introduces a hazard caused by the transit time of A relay contacts in passing from relay combination 9 to 10. Therefore the original expression will be used for relay D . The control path is shown in Fig. 3c.

For the E relay which operates and locks only once in the cycle

$$g(E) = (A' + B' + C' + D)$$

$$h(E) = E$$

and

$$\begin{aligned} f(E) &= (A' + B' + C' + D)(E + E) \\ &\quad (A' + B' + C' + D)E \end{aligned} \quad (4a)$$

This control path is shown in Fig. 3d.

Apart from the problem of developing the required contact network, the practical problem of what operating power must be given to the relays in order to meet speed requirements must be dealt with. Since the use of low resistance windings in series with protective external resistors is called for to obtain the speed required, it appears that the use of two windings per relay might prove advantageous. By operating on the low resistance winding while locking on the high resistance winding, the current drain may be reduced (thereby saving a fuse) and furthermore some code reduction may be made possible as shown later. If double windings are used, two of the external series resistors may be eliminated by combining the control network so as to make certain that only one of the low resistance windings on the C , D , or E relays is energized at any one time. This would permit the use of one common external resistance with the aforementioned relays instead of three.

Keeping these practical considerations in mind, further savings may be made by combining the control circuits as shown in Fig. 3e. Although there is in this circuit a possibility of contact stagger on the A relay contacts causing the C and D low resistance windings to be energized at the same time, this will not be harmful since, when the stagger occurs, both relays are firmly locked operated by their high resistance holding windings.

TRANSLATING CIRCUIT

The translating circuit is particularly adaptable to switching algebra manipulation. Table I shows the combinations which prevail at the end

of each pulse and the necessary "code" leads that must be grounded at these times. Reference to the block diagram of Fig. 1 shows that the output of the translator is not activated until the slow release *RA* relay releases after the last pulse of a digit has been received. Therefore the *A* relay can be eliminated from these combinations since at the end of every pulse the *A* and *B* relays are either both operated or both released and hence only one is needed to indicate the condition of both. Table III lists the numerous combinations which must close a ground path through to each of the five code leads and the *Z0* lead. At the conclusion of the algebraic manipulation, the *A* and *B* contacts may be redistributed evenly since they perform interchangeable functions in translation.

The objective in the design of the translating circuit is to obtain the most economical contact network subject to a spring distribution that

TABLE III
TRANSLATION

Output Lead Grounded	Counting Relays				Decimal Pulse
	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	
0	0	1	1	1	1
	1	0	1	1	2
	1	0	0	1	4
	0	1	0	0	7
	0	1	1	0	11
	1	1	1	0	12
1	0	1	1	1	1
	0	0	0	1	3
	0	1	0	1	5
	1	0	0	0	8
2	1	0	1	1	2
	0	0	0	1	3
	1	1	0	0	6
	0	0	1	0	9
4	1	0	0	1	4
	0	1	0	1	5
	1	1	0	0	6
	1	0	1	0	10
7	0	1	0	0	7
	1	0	0	0	8
	0	0	1	0	9
	1	0	1	0	10
<i>Z0</i>	1	0	1	0	10

Inconsequential combinations that may be used for simplification

0	0	1	1
1	1	0	1
0	0	0	0
1	1	1	1

fits in with the control contact network. Again, this is a multi-terminal network problem and the procedure is to design two-terminal networks that combine most readily. Since it is impractical to illustrate all the repeated trials that led to the final design, each network will be designed separately with the understanding that some of the steps are imposed by the form of all networks viewed collectively.

The procedure adopted for developing the "0" lead network is as follows. First set up the miniature table repeating the portion of Table III that corresponds to the "0" lead. These parallel combinations should then be manipulated algebraically to obtain the greatest simplification possible. It is rather easy to apply some of the algebraic rules by observing the condition of the relay in the several combinations in the table. A simple "shorthand" rule to follow is: if in the table of combinations describing a particular two terminal network, all possible combinations of certain relays appear in conjunction with a single combination of other relays, the network contacts on the former relays may be neglected. In other words when 2^n different combinations of any number of variables m , are identical in all but n columns, contacts on the corresponding n relays are not required. This procedure is carried out below.

B	C	D	E	
0	1	1	1	— (B + C' + D')
1	0	1	1	— (B' + C + E')
1	0	0	1	— (B' + C + E')
0	1	0	0	— (B + C' + E)
0	1	1	0	— (B + C' + E)
1	1	1	0	— (C' + D' + E)

Thus we have the following algebraic expression for the "0" lead, which can be simplified as shown.

$$(B + C' + D')(B' + C + E')(B + C' + E)(C' + D' + E) \\ [C' + (B + D')(B + E)(D' + E)](B' + C + E') \quad (3b)$$

$$[C' + (E + BD')(B + D')](B' + C + E') \quad (3b)$$

This is shown on Fig. 4a. A somewhat different manipulation of the equation permits placing the network in the bridge form of Fig. 4b. The algebraic equation, given below, can easily be shown to be the equivalent of the original.

$$[E + C' + B(B' + D')](B' + C + E')(B + C' + D')$$

In certain cases the use of theorem (14b), normally employed to reduce the contacts of a particular relay to a single make and break, can produce simplifications difficult to accomplish otherwise. This is shown below, with the theorem applied with respect to relay E since E tended otherwise to be heavily loaded.

$$(B + C' + D')(B' + C + E')(B + C' + E)(C' + D' + E)$$

$$[E + (B + C' + D')(B' + C + 1)(B + C' + 0)(C' + D' + 0)]$$

$$[E' + (B + C' + D')(B' + C + 0)(B + C' + 1)(C' + D' + 1)]$$

$$(14b)$$

$$(E + C' + BD')[E' + (B' + C)(B + C' + D')] \quad (9a, 10a, 3b, 9b, 5b)$$

By modifying the first factor of the final expression in accordance with theorem 11a, this equation can be put in bridge form as shown on Fig. 4c.

$$[E + C' + B(B' + D')][E' + (B' + C)(B + C' + D')]$$

The above equation uses the same contacts as the previous expression, and although the right hand member is in a slightly different form, the expression is equivalent to the one obtained earlier.

When it is known that output conditions are inconsequential for some relay combinations, these inconsequential relay combinations may be combined with valid combinations to eliminate contacts in the network. Inconsequential means that the output during these particular combinations does not affect the proper functioning of the circuit. Four such combinations are listed in Table III. Only those inconsequential combinations which will combine readily with the actual combinations, thereby resulting in a reduction in the number of contacts, are to be used. Although the use of all the all-relays-released condition may be helpful in certain cases, it will not be used in the circuit under consideration since its use makes the requirement that no tie shall exist between output leads until the second pulse is received hard to meet.

With this in mind the "0" lead network is again examined. Note the use of another "shorthand" rule which states that if a part of the 2^n possible combinations is used in closing a path, the negative of the unused part of the 2^n possible combinations is equivalent to the original combinations. Thus if in the case at hand three of the possible four combinations of the B and C relays occur in series with the same combination of the D and E relays, the expression used is that for the series path of the D and E relays plus the negative of the missing combination of the B

and C relays. In the following tabulation the combinations below the horizontal line are inconsequential.

B	C	D	E	
0	1	1	1	(C' + D' + BE)
1	0	1	1	
1	0	0	1	(C + E' + B'D')
0	1	0	0	
0	1	1	0	(B + D + E)
1	1	1	0	
<hr/>				
0	0	1	1	(B + D + E)
1	1	0	1	
0	0	0	0	

The expression becomes:

$$(C + E' + B'D')(C' + D' + BE)(B + D + E)$$

$$[E + (C + 1 + B'D')(C' + D' + B0)(B + D + 0)]$$

$$[E' + (C + 0 + B'D')(C' + D' + B1)(B + D + 1)]$$

(14b on E)

$$[E + (C' + D')(B + D)][E' + (C + B'D')(C' + D' + B)]$$

(9a, 9b, 10a, 10b)

$$[E + (C' + D')(B + D)][E' + CD' + CB + B'D'] \quad (8b, 9a, 4b, 5a)$$

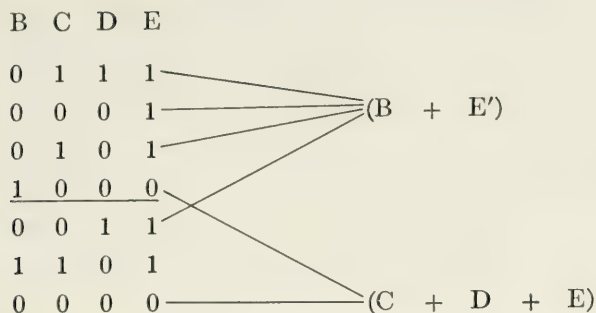
$$[E + (C' + D')(B + D)][E' + BC + B'D'] \quad (12b)$$

$$[E + (C' + D')(B + D)][E' + (B + D')(B' + C)] \quad (13)$$

Fig. 4d shows the schematic of the above expression. It is possible to put this in a bridge form without other changes because of the manner in which the front and back contacts of *D* are related to the other contacts. Comparison of all the circuits of Fig. 4 indicates that they all use the same number of contacts although final decision should be postponed until all the output circuits are obtained and the ease of combination of the different circuits can be compared.

The procedure for determining the remaining code leads is carried out on the following pages.

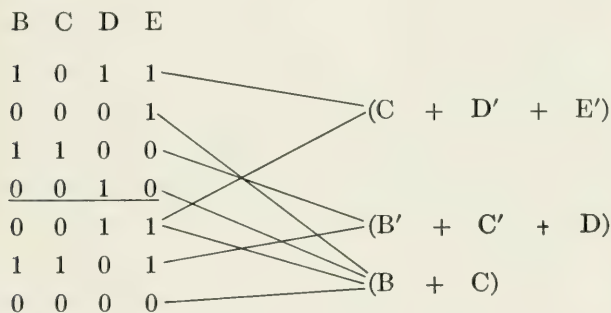
"1" lead—



resulting in $[E + C + D][E' + B]$ which is shown on Fig. 5a. It will later be found advantageous, in combining, to include the B' term in the first factor, giving the expression:

$$(E + B' + C + D)(E' + B) \quad \text{shown on Fig. 5b}$$

"2" lead—



hence

$$(C + D' + E')(B' + C' + D)(B + C)$$

or

$$[C + B(D' + E')][C' + B' + D] \quad (3b)$$

For later ease in combining, this is changed to:

$$[C + B(B' + D' + E')][C' + B' + D] \quad (11a)$$

shown on Fig. 5c.

" γ " lead—

B	C	D	E	
1	0	0	1	(D + E' + B'C')
0	1	0	1	
1	1	0	0	(B' + C + D' + E)
<u>1</u>	<u>0</u>	<u>1</u>	<u>0</u>	
0	0	1	1	(B' + C' + D)
1	1	0	1	
0	0	0	0	

hence

$$(D + E' + B'C')(B' + C' + D)(B' + C + D' + E) \\ [D + (B' + C')(B'C' + E)][D' + B' + C + E] \quad (3b)$$

which is shown on Fig. 5d.

" γ " lead—

B	C	D	E	
0	1	0	0	(B + D + E)
1	0	0	0	
0	0	1	0	(C + E)
<u>1</u>	<u>0</u>	<u>1</u>	<u>0</u>	
0	0	1	1	(B + D + E)
1	1	0	1	
0	0	0	0	

hence

$$(B + D + E)(C + E)$$

or

$$E + C(B + D) \quad (3b)$$

which is shown on Fig. 5e.

"Z0" lead—

B	C	D	E
1	0	1	0

hence one has $(B' + C + D' + E)$ which is shown in Fig. 5f.

The final contact savings are achieved by combining the various output paths. The combined translation circuit that appears to be as reduced as possible is shown in Fig. 6. Note that certain forms of the individual output paths combine more readily than others. For example Fig. 4c and 5b combine more readily with the remaining paths than Fig. 4b and 5a. Note also that sometimes it is not the most reduced form of the individual output paths that permits efficient combining. This is

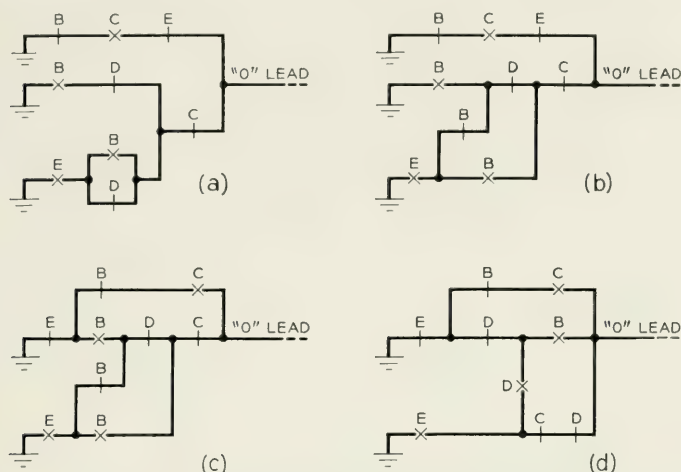


Fig. 4—The "0" lead of the translating circuit.

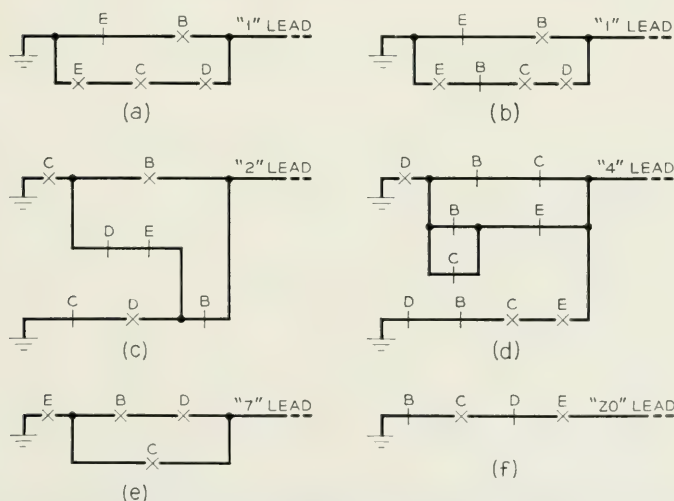


Fig. 5—The "1, 2, 4, 7, and Z0" leads of the translating circuit.

exemplified by the use of Fig. 5b rather than Fig. 5a. Although various forms of all of the output leads were tested for efficient combination only the form used is shown for outputs other than the "0" and "1" leads.

It is essential to scrutinize the final network for possible sneak paths. Sometimes to avoid these sneak paths it is necessary to add one contact on one relay to allow savings on others. Here again the inspection techniques go hand in hand with switching algebra and the need for both is obvious. The algebra obtains the various forms which are capable of

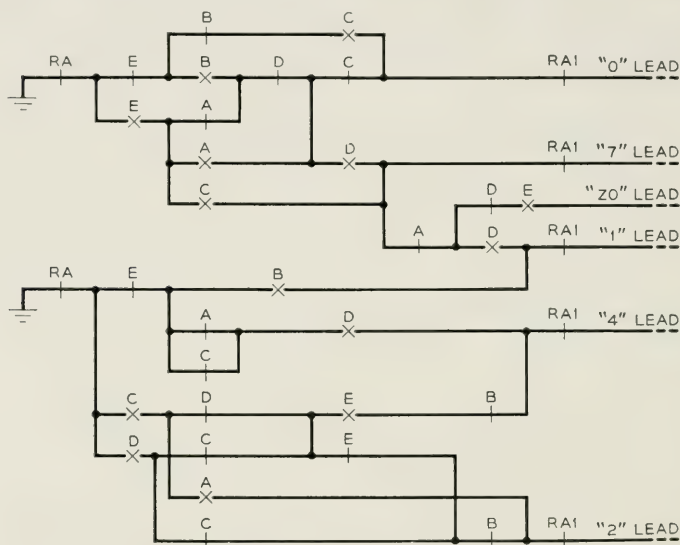


Fig. 6—Combined translating circuit.

different degrees of combination very quickly and efficiently. The inspection method is then necessary for the actual combination of these forms.

The additional *RA* relay contact is necessary to assist in avoiding interconnections between the output leads until after the second pulse is received. The final assignments of either *A* or *B* relay contacts are chosen to equalize the load on these relays.

THE COMPLETED CIRCUIT

The final form of the counting and translating circuit is shown on Fig. 7. The relays are all double wound to gain the benefits of current drain reduction. One additional advantage of using double windings is the relay code reduction made possible since now only two codes are necessary. One code serves the *A* relay and one other code serves the

B , C , D and E relays. In comparison to this total of five relays and two codes the circuit in present use in the latest crossbar system requires ten relays and seven codes.

AN ALTERNATE DESIGN OF THE PULSE DIVIDER

To illustrate the application of algebra where the apparatus contemplated puts less premium on contact minimization but more on standardization and winding minimization, certain modifications of the proposed circuit are considered.

In the event that new apparatus developments make possible the construction of relays that meet the necessary speed requirements even though winding impedance is increased, it appears possible (if the pulse divider is redesigned) to use only one code having a single winding for all five relays. The use of added contacts might be allowable if the new type of relay carries more springs than the present relay.

The redesign of the pulse divider to use single windings can be accomplished by manipulation of the basic algebraic expressions derived earlier for the pulse divider.

Thus for the A relay

$$\begin{aligned} f(A) &= (L' + B')(A + LB') \\ &= [L' + B'] [A + B'(L + B)] \end{aligned} \quad (11a)$$

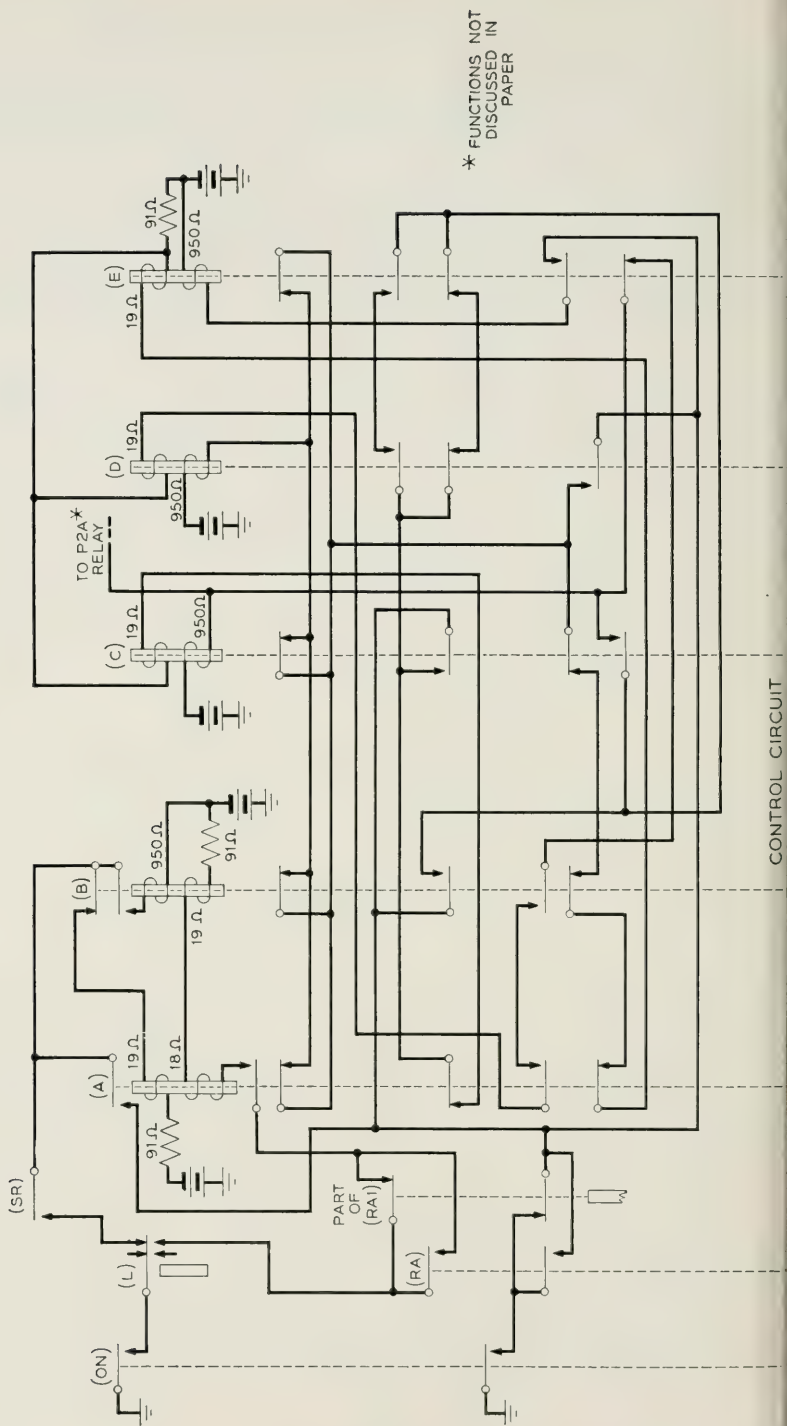
By attempting to manipulate the B relay control circuit into the same form, one obtains

$$\begin{aligned} f(B) &= (L + A)(B + L'A) \\ &= (L + A)(L' + B)(A + B) \\ &= [L' + B][A + BL] \\ &= [L' + B][A + B(L + B')] \end{aligned} \quad \begin{aligned} (3b) \\ (3b) \\ (11a) \end{aligned}$$

The schematics represented by the above algebraic expressions are shown in Fig. 8a and 8b. The circuit of Fig. 8c is obtained by combining the first two circuits so that only a single transfer is needed on the L relay. Note however that it is necessary to make the lower two B transfers have continuity action to insure proper functioning. Fig. 8d shows the pulse divider drawn in conventional form.

CONCLUSION

As far as is known, the dial pulse counting and translating circuit described herein requires fewer relays than any other circuit with similar



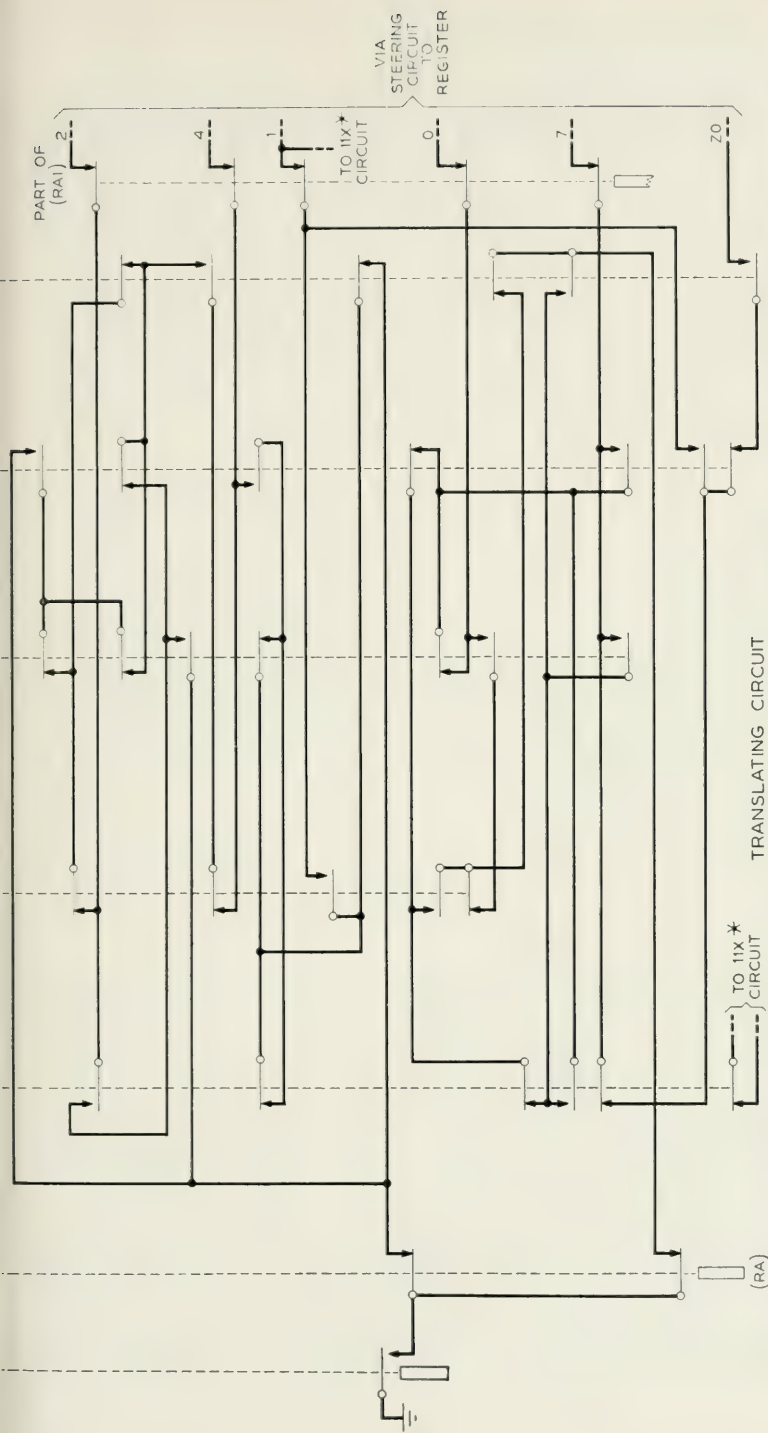


Fig. 7—The complete all-relay dial-pulse counting and translating circuit employing the minimum number of relays.

functions at present employed in Bell System standard switching equipment. The previous dial pulse counter used in the latest crossbar system required a total of ten relays. Thus the present design represents a considerable saving in cost and space. To a certain extent this result can be ascribed to the use of switching algebra during the circuit development.

Relay circuits designed on the basis of utilizing a large proportion of the possible combinations permitted by the component relays usually require heavy spring pile-ups. Since general purpose relays are limited in the number of springs which they can carry, this type of circuit usually

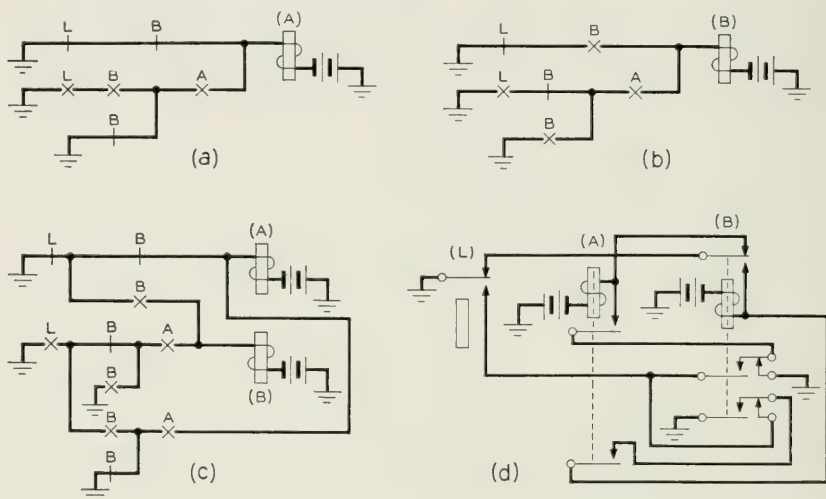


Fig. 8

entails considerable design effort to make most effective use of the available springs. Application of switching algebra to this aspect of the design problem can often provide crucial assistance.

It is recognized that switching algebra, in its present state of development, does not permit complete mathematical statement and manipulation of multi-terminal networks as represented by the counting and translating circuits. It does provide, however, facilities in manipulating two-terminal networks into a variety of forms from which can be selected those that combine most readily. This can result not only in a saving of time, but also in improved circuits which might not be realized by other design techniques. Unfortunately the algebra in its present state does not indicate when the optimum circuit has been attained. To some extent this is caused by apparatus or circuit considerations to which, since it is

concerned solely with contact networks, the algebra does not apply. Thus, there is still considerable room left for the ingenuity and judgment of the switching circuit designer.

As a result of the experience in designing the dial pulse counter and translator, certain observations on the use of the algebra are believed to be valid. Although switching algebra may be used in the design of the simplest circuits, the most noticeable benefits are obtained by the application of the algebra to the design of those circuits in which the control and output paths are complex and interrelated. The particular minimum relay counting and translating circuit under discussion is an excellent example of this type of circuit. A secondary advantage of the algebra is its compact notation and its value as an efficient circuit "bookkeeping" method.

ACKNOWLEDGMENTS

The author is indebted to Joseph Michal who made invaluable contributions as co-designer of this circuit and also verified all the algebraic manipulations contained in this paper. The author also wishes to acknowledge the suggestion made by G. R. Frost as to the possibility of using the bridge network in the control circuit. This work was carried on under the supervision of L. J. Stacy and F. K. Low whose many valuable suggestions were incorporated into this development.

Interaction of Polymers and Mechanical Waves

BY W. O. BAKER AND J. H. HEISS

(Manuscript received October 19, 1951)

New techniques of Mason, McSkimin, Hopkins and co-workers for generation of shear waves over the frequency range 2×10^2 to 2.4×10^7 cps have been used to study mechanical properties of chain polymers. Polymer solids, melts and dilute solutions, representing the main states in which plastics and rubbers are fabricated or used, were explored to find the characteristic relaxation times, rigidities and viscosities of various chemical structures. Polyisobutylene, hevea rubber, polydimethyl siloxane, vinyl chloride-acetate copolymers and plasticized nitrocellulose were compared with polyethylene and polyamides as examples of the range of solid properties encountered.

As melts, several polyisobutylenes, polybutadiene, polypropylene, polypropylene sebacate and poly- α -methyl styrene were investigated as models for varying degrees of chain substitution. Chain rigidity in, for instance, polyisobutylene, seemed to reflect visco-elastic over-all configurational changes up through the kilocycle range, but nearest neighbor interactions took over in the megacycle region, leading to moduli of 10^9 dynes/cm² even for syrupy fluids.

In dilute solution, polyisobutylene, polystyrene, natural rubber and polybutadiene microgel exhibited characteristic dynamic viscosities and rigidities depending linearly on concentration. Presumably, this reflects mechanical properties of isolated chains. Some possible models were suggested for the frequency dependence of such properties.

INTRODUCTION

The "equilibrium" mechanics of polymers, the giant molecules of plastics and rubbers, have been quite elegantly developed in the range of high strains ("kinetic theory" of elasticity—Meyer,¹ *et al.*). However, the molecular displacements as these strains, and, indeed, much smaller ones, occur, are little understood.² Nevertheless, it is essential to know about detailed motions in connecting chemical structure with physical properties. Only in this way can there be obtained from the chemical industry compositions which will serve properly in telephone apparatus.

Other studies have treated one way of getting at these mechanisms by relating stress relaxation, creep, viscosity, etc. to a distribution of molecular relaxation times (and energy barriers), as originated by Kuhn.^{3, 4} Another approach is to strain polymers with periodic waves over a very wide spectrum of wavelengths, eventually going to frequencies comparable with those of the thermal vibrations of significant groups or segments in the macromolecules. The resulting dispersion or resonance phenomena can then be examined. Hence a mechanical radiation field can interact with the masses of elementary structural units, as the usual electromagnetic field interacts with atomic and group charges. In general, direct interpretations of this kind must be done with shear waves, and, at least, not *only* with longitudinal or ultrasonic waves.

This kind of study is now proceeding using waves generated and followed by piezoelectric crystals connected in as actual electromechanical circuit elements (A. M. Nicolson, 1919). Recent schemes of Mason and co-workers cover the frequency range from 10×10^3 to 60×10^6 cps, as reported in the paper by Mason and McSkimin in the last issue, while a tuning fork method used by I. L. Hopkins has been applied to "soft" polymers (rubbers) over the range 10^2 to 10^4 cps (the general range of J. D. Ferry's work at Wisconsin on concentrated polymer solutions).

The relation of these studies to the scientific and technical exploitation of plastics and rubbers is in knowing what a particular chemical composition does to strength, stiffness, ease of molding, impact toughness, etc. That is, are there qualities of the interaction of saturated aliphatic groups that make polyethylene or polyisobutylene have some glass-like as well as liquid-like, or rubbery, nature even at room temperature? If so, conditions causing brittle failures must be watched for. How is the storage of molecular strains in injection molded plastics reduced by increasing molding temperature (when the kinetic theory stiffness per chain actually increases)? These and many similar problems may be generalized under the headings below; in each case the chemical structure of the macromolecule appears to be reflected in relaxation times which combine in different ways to give flow or rigidity, toughness or brittleness.

Extrusion and Molding

Non-Newtonian flow leading to "frozen-in" stresses, subsequent distortion and irregular shapes of plastics⁵ and rubbers,⁶ implies energy

storage in the sheared molecules. The dynamic shear studies will confirm this. Also dispersion of carbon black and other pigments is restrained by elastic qualities of "liquid" polymers (i.e., instead of "mixing", compounds just microscopically deform and later re-form.) Likewise, the efficiency of compounding⁷ and extrusion⁸ depend on how quickly the molecules relax after straining.

Impact Strength, Brittleness and Tenacity

Toughness, mechanical shock resistance, ultimate elongation and strength reflect the facility with which the polymer molecules can be displaced without breaking the piece. Thus, they accommodate to the stress by motions presumably similar to those described above. (The situation is complicated when crystallites are also displaced.⁹) In any case, time sensitivity in the range 10^{-5} sec upward exists.^{10, 11} The discussion by Morey¹¹ is a valuable survey of these ideas, and explicitly notes the significance of multiple relaxation processes on damping of shock waves. Evidence of the relation of simple changes in chemical structure to the principle relaxation times effective in these physical properties of plastics and rubbers is thus another part of the dynamics studies. The "brittle point", or volume-temperature transition of amorphous polymers,^{12, 13} apparently reflects directly the correspondence of the time of experiment with dominant relaxation time of the polymer.^{14, 15} A few measurements on plasticized polymethyl methacrylate (from which, however, no actual rigidities were calculated) indeed indicate abrupt stiffening as a function of frequency at a given temperature.¹⁶ However, the changes measured were too small and indefinite to indicate any particular molecular relaxation. Other work¹⁷ with plasticized polymers is nevertheless concordant with the current findings that molecular relaxations and not long range order determine embrittlement. The converse of this is, of course, that as some "transition" is approached, hysteresis, heat build up, flex cracking and fatigue are greatest.

Creep, Stress Relaxation and Recovery

Even these "long time" qualities of plastics, such as found in cold flow, apparently result from integrated displacements of rapidly oscillating segments of the chains. A most interesting analysis of stress relaxation in rubbers employs Kuhn's suggestion of a particular distribution of relaxation times.¹⁸ The present point is that, again, these relaxation times reflect processes which should appear directly in reaction of the polymer with high frequency shear waves.

From these aspects above the current results of dynamics studies will be reviewed.

POLYMER SOLIDS: OVER-ALL MECHANICS

Solid polymers will denote rubbers and plastics in the state in which they are technically used. This is usually their most complex form, with inter- and intra-molecular factors undistinguished. Thus, separation and identification of the main relaxation processes are difficult or impossible. However, it is interesting to consider typical values of modulus and viscosity as related to chemical structure, in the range of frequencies corresponding to extrusion rates, and stresses in actual use.

These values of dynamic modulus and viscosity are distinct from the usual quantities in the literature. The usual expressions are for longitudinal (sound) waves, and give dynamic Young's modulus¹⁹

$$E^* = E_1 - iE_2$$

E_2 measures the out of phase part of the force-displacement relation, and $E_2 = \omega \cdot$ ("effective viscosity coefficient"). Now, the general elastic constants are $\lambda + 2\mu$, with λ = Lamé's constant and μ = shear modulus. Here,

$$\lambda + 2\mu = K + \frac{4}{3}\mu,$$

with K = bulk modulus. Alternately,

$$E_1 = \frac{3K}{\lambda + \mu} \mu = \frac{3\lambda + 2\mu}{\lambda + \mu} \mu.$$

However, in general the present results lead to the simpler shear modulus μ . Further the energy losses studied are expressible directly as the usual shear viscosity

$$\mu' = \eta.$$

Previous comprehensive studies of the dynamics of rubbers over significant frequency ranges have yielded loss factors either written as E_2/E_1 (see above),¹⁹ or as a function of the shear viscosity based on Stoke's assumption that the compressional (dilatational) viscosity is zero.²⁰ But as Nolle¹⁹ and Ivey, Mrowca and Guth²⁰ clearly recognize, recent work has strongly manifested the presence of compressional viscosity in simple liquids²¹ as well as polymeric ones.^{22, 23} Hence, the present understanding relating molecular structure to viscosity, plasticity and visco-elasticity is unsuitable for interpreting mechanical wave motion more complex than in shear, unless shear constants are also known.

This sums up to mean that the *chemical* interpretation of basic polymer mechanics requires shear wave measurements. Nevertheless, fascinating evidence of the existence of fine-structure relaxations in polymer solid has come from longitudinal wave investigations.^{19, 20, 24, 25, 26} Also, the pioneering shear wave studies of Ferry and collaborators^{27, 28} on concentrated solutions of polymers have suggested intrinsic relaxations of the chain molecules in a highly plasticized "semi-solid" state.

The more simplified findings cited below will be seen to unify approaches in this field. Comment must first be made, however, on formulation of experimental results in dynamics of polymers.

Expression of Dynamic Properties

Alternate and equivalent expressions have been thoroughly surveyed;²⁹ all represent combinations of either Maxwell (series) springs and pistons (elasticity and viscosity) or Voigt (parallel) springs and pistons. Obviously, there is no physical separation of elastic and viscous elements in a polymer molecule, so the irrelevance of the *detail* of the model need not be emphasized. However, the models lead to convenient formulation of *relaxation times* which dielectric studies, in particular, have shown have clear connections with chemical structure. In this chapter, sometimes one and sometimes the other model, or combination, will be used, with the symbols shown on the next page.

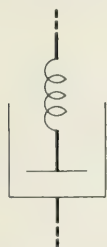
Other symbols are sometimes used,³⁰ but should be easily identified in terms of the above.

Rubbers and Soft Plastics

In Table I, the shear moduli of rigidity, μ , and of viscosity, μ' , are shown as calculated for the Kelvin-Voigt model, for polymers having the indicated units of structure. The frequencies are from a few hundred to a few thousand cycles, hence, in the range of much technical use, (flexing of tires ~ 300 cps) and rates of shear during processing.^{31, 32} Data are from a general study by I. L. Hopkins³³ of the Bell Laboratories, based on a tuning fork transducer introduced by Rorden and Grieco.³⁴ The strains employed were always small, in the range 0.3 to 1.5 per cent; μ and μ' were essentially independent of strain, except for some loaded rubber stocks. The μ values clearly trace the magnitudes to be expected in going from the most typical rubber (hevea) to the semi-rigid plastics (vinyl chloride-acetate copolymer and plasticized cellulose nitrate). As anticipated from steady-stress observations the "plastics" have $\mu > 10^7$ dynes/cm². Increase of μ with frequency is also greater as

the "plastics" range is approached; a relaxation region is implied. Figs. 1 to 4 show the dispersion of rigidity with frequency in more detail. Especially striking in Figs. 1 and 2 is the small temperature dependence (at least between 27° and 66°C) of μ . Because of experimental uncertainty, μ cannot be said to be actually higher at the higher temperatures in accord with straight kinetic theory, but at least it is strongly tending that way, as also noted for lower frequencies studies on natural rubber.¹⁹ Nothing like this appears for the plastics; in plasticized nitrocellulose the 100-cycle rigidity decreases 10-fold from 27° to 66°C. This is, then, the second general dynamic quality which reflects the low van der Waals' (dipole, dispersion and induction) forces in hevea rubber and polydimethyl siloxane, as well as their intrachain flexibility. Interchain forces in polyisobutylene (Butyl rubber) are low too, but barriers to flexibility because of sterically hindered-CH₃ groups come in. Table I and Fig. 3

MAXWELL



σ = strain
 S = stress
 t = time
 τ = relaxation time
 τ' = retardation time
 μ = G = modulus
 μ' = η = viscosity

$$\frac{d\sigma}{dt} = \frac{1}{\mu} \frac{dS}{dt} + \frac{S}{\eta}$$

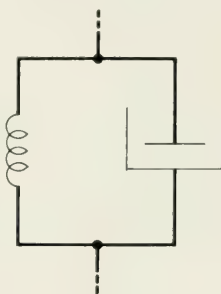
$$S = S_0 e^{-\frac{\mu t}{\eta}} = S_0 e^{-\frac{t}{\tau}}$$

$$\tau = \frac{\eta}{\mu}$$

For const. S , $\frac{d\sigma}{dt} = \frac{S}{\eta}$ or $\eta = \frac{S}{\frac{d\sigma}{dt}}$

There is same stress on each element; the total strain = sum of single strains.

KELVIN-VOIGT



$$\frac{d\sigma}{dt} = \frac{S}{\eta} - \frac{\mu}{\eta} \sigma$$

$$\sigma = \frac{S}{\mu} \left(1 - e^{-\frac{\mu t}{\eta}} \right) = \sigma_0 e^{-\frac{t}{\tau'}}$$

$$\tau' = \frac{\eta}{\mu}$$

There is same strain in each element; the total stress = sum of single stresses.

TABLE I

Polymer Unit	Shear Modulus, μ , dynes/cm ² 27°C		Shear Viscosity Poises, μ' , 27°C	
	100 cycles	5000 cycles	100 cycles	5000 cycles
Hevea rubber	3×10^6	5.5×10^6	350	40
$\begin{array}{c} \text{CH}_3 \\ \\ -\text{CH}_2-\text{C}=\text{CH}-\text{CH}_2- \end{array}$				
Polydimethyl siloxane	0.7×10^6	1×10^6	300	30
$\begin{array}{c} \text{CH}_3 \\ \\ -\text{O}-\text{Si}- \\ \\ \text{CH}_3 \end{array}$				
Polyisobutylene	5×10^6	30×10^6	8,000	1,500
$\begin{array}{c} \text{CH}_3 \\ \\ -\text{CH}_2-\text{C}- \\ \\ \text{CH}_3 \end{array}$				
Polyvinyl chloride (~92%)-acetate (~8%) plasticized by ~36% di- octylphthalate	13×10^6	80×10^6	25,000	2,000
$\begin{array}{c} -\text{CH}_2-\text{CH}- \\ \\ \text{Cl} \end{array} \quad \text{and}$				
$\begin{array}{c} -\text{CH}_2-\text{CH}- \\ \quad \text{O} \\ \text{O}-\text{C}=\text{CH}_3 \end{array}$				
Cellulose nitrate	60×10^6	250×10^6	80,000	4,500
$\begin{array}{c} \text{CH}_2\text{ONO}_2 \\ / \quad \backslash \\ \text{CH}-\text{O} \quad \text{O} \\ / \quad \backslash \quad / \quad \backslash \\ -\text{C} \quad \text{CH}-\text{O}- \\ \quad \\ \text{CH}-\text{CH} \\ \quad \\ \text{ONO}_2 \quad \text{ONO}_2 \end{array}$				
and ~25 wt. % Camphor plasti- cizer.				

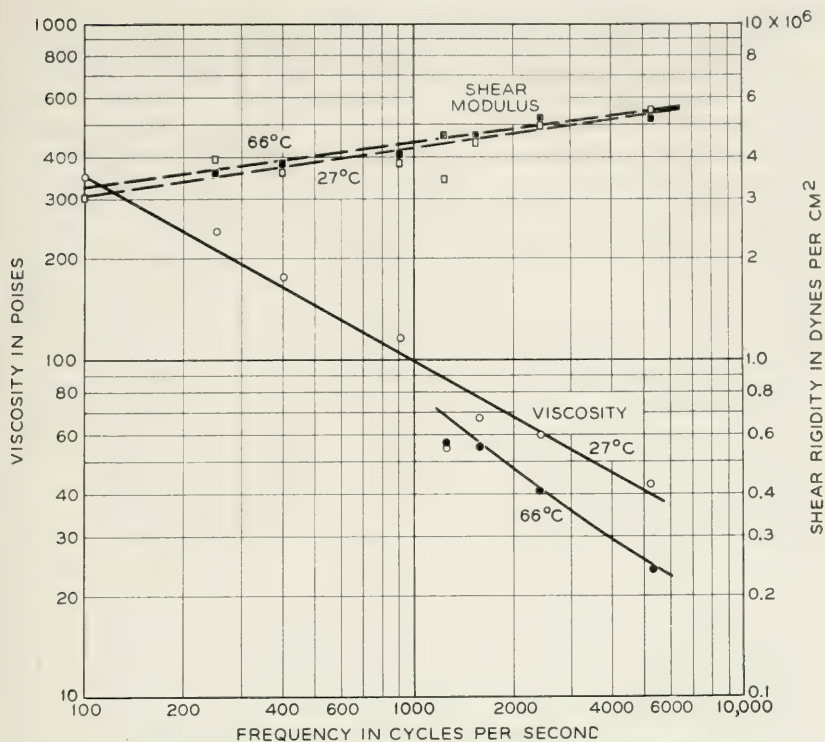


Fig. 1—Viscosity and shear modulus of hevea rubber (cross-linked).

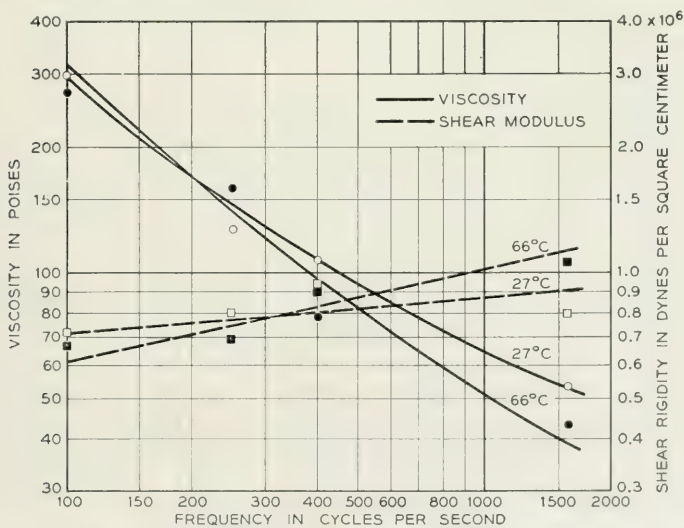


Fig. 2—Viscosity and shear modulus of polydimethyl siloxane (cross-linked).

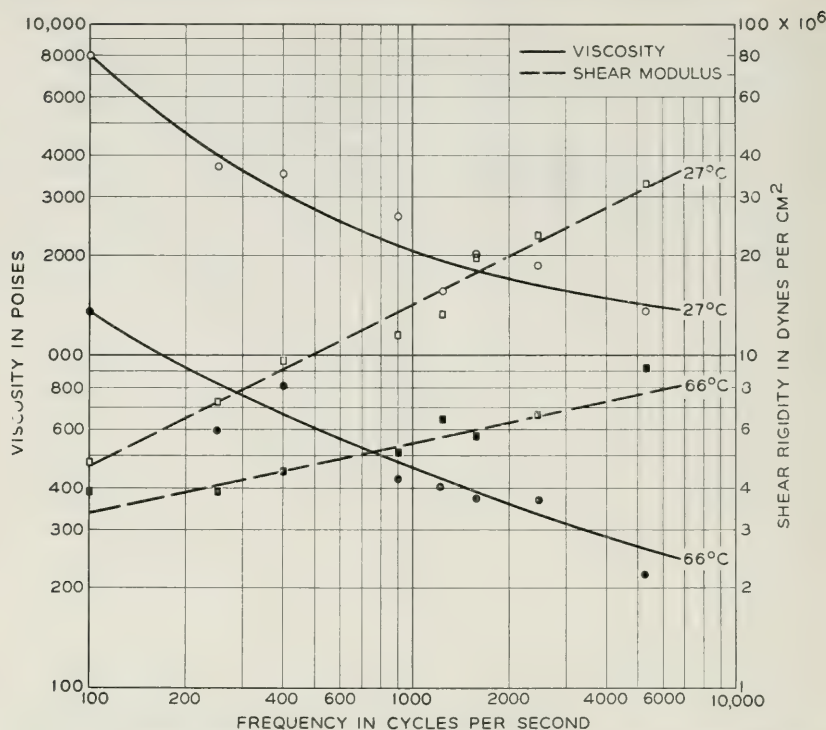


Fig. 3—Viscosity and shear modulus of polyisobutylene (cross-linked Butyl).

emphasize that thinking about the mechanics of a particular chemical structure must include the spatial relationships of groups within the chains, as well as between them.

The dynamic viscosities in Table I are also in accord with the sequence of structures. Their frequency dispersion again connotes varying relaxation processes. Natural rubber's low inner friction, for both compressional³⁵ and shear waves³⁶ is famous in its low hysteresis heating. (This unique property is geopolitically crucial, because adequate truck and bus tires cannot yet be made of any other rubber.) Indeed, it is striking that at 100 cycles, a piece of gum rubber has a local viscosity of only 350 poises. The silicone rubber gum also has high elastic efficiency, and its temperature coefficient of viscosity is very low (see Fig. 2), like the thermal coefficients for familiar silicone liquids. It is exciting to speculate in Figs. 1 and 2, whether more precise measurements which Hopkins is now undertaking will confirm the apparently *negative* temperature coefficients of viscosity at some frequencies. "Kinetic theory

viscosity" arising from transfer of momentum among thermally agitated chain segments, does not seem to have been considered in the theory of perfect rubbers. As in gases, it would require an increase of viscosity with temperature.

In polyisobutylene, however, the dynamic viscosity leaps upward in both magnitude and temperature dependence. It should be emphasized that this is, again, for a cross-linked (Butyl) gum—an infinite network like the hevea gum, with presumably infinite macroscopic viscosity. The striking thing is that this internal viscosity is not greatly dependent on the network, at the degrees of "cure" used in rubber technology. For instance, recent studies over the frequency range 20–600 cycles, on high molecular weight, $\bar{M}_\eta = 1.2 \times 10^6$, linear polyisobutylene,³⁷ give, at 25°C and 100 cycles, $\mu' = 4800$ poises, although the steady flow viscosity of this polymer at this temperature is greater than 3×10^9 poises.³⁸ Then, the infinite network ($\eta_{\text{steady flow}} \rightarrow \infty$) Butyl polymer of Fig. 3 has at 27°C and 100 cycles $\mu' = 8000$ poises. At 1000 cycles agreement appears to be about the same, and is tolerable considering the several

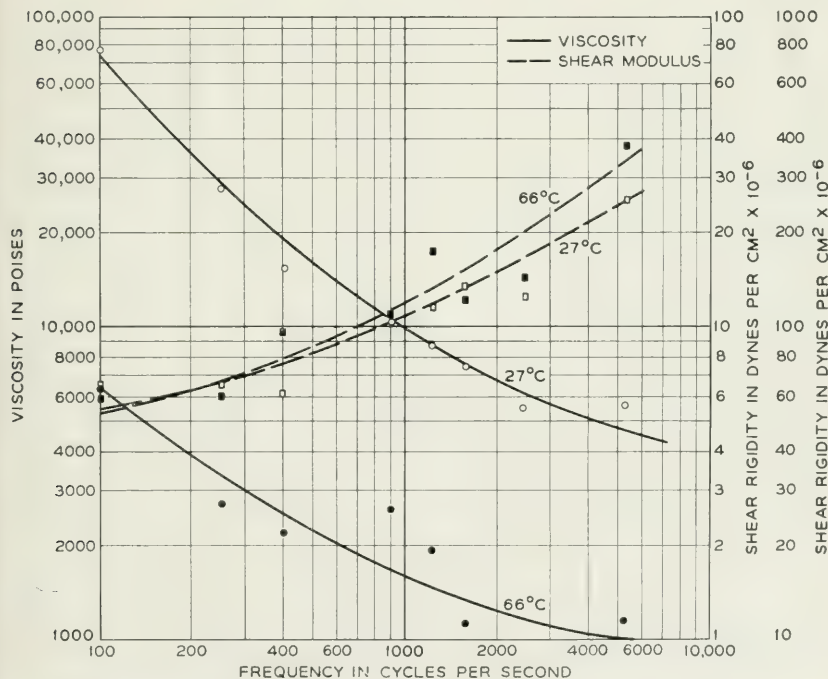


Fig. 4—Viscosity and shear modulus of plasticized cellulose nitrate.

per cent of compounding ingredients in the Butyl gum, and possibility of a small fraction of low molecular weight uncured polymer in it.

Also, wide variations in the degree of cure of Butyl gums were studied without large changes in μ' .³³ In this regard, the particular sample of Fig. 3 had an equilibrium swelling ratio (= volume swollen polymer in cyclohexane at 25°C/volume insoluble part of dry vulcanizate) of 4.84. This indicates an M_c value (average molecular weight between cross-links) of $<20,000$.³⁹ Actually many of the dynamic properties can probably be found in individual chain units or segments even smaller than this. This is a significant point in engineering applications where plastics may be cured to reduce creep but where it is desired to retain typical "chain" properties to increase impact toughness. That is, usually some optimum condition for this compromise can be found. The later section on liquids will suggest that physical properties typically associated with chain polymers can indeed reside in even shorter chain sections than the M_c 's observed in usual gum rubbers.

Filled Polymers

Marked effects of carbon black and other pigments are of course familiar in both steady and alternating mechanics of rubbers.^{19, 24, 35, 36} Brief comment on their influence on dynamic shear properties and thus relaxation mechanisms involved may be directed toward plastics, also, however. Thus, technologically it would be desirable to load thermoplastics with considerable volumes of "inert" fillers, just as is done with rubber. But, almost invariably strength and toughness decline, instead of improving, as in the rubber case. A reason for this appears in investigations by Hopkins when carbon black (a standard type of reinforcing black) was added to Butyl rubber of the sort described in Table I. It is that stiffness seems to rise more rapidly than internal viscosity—i.e., a given strain results in proportionately higher stress than the accompanying internal viscosity provides means for dissipating the stress (as on impact). Hence, the brittleness which fillers normally engender in thermoplastics may represent this change in μ vs. μ' balance. Table II illus-

TABLE II

Wt. Per Cent of Carbon Black in Butyl Stock	Shear Modulus, μ , dynes/cm ² at 27°C		Shear Viscosity, μ' , Poises, at 27°C	
	100 cycles	5000 cycles	100 cycles	5000 cycles
15.2	8×10^6	60×10^6	11,000	2,000
28.6	45×10^6	150×10^6	35,000	3,000

trates some values for Butyl compounds. The swelling ratio (SR) for the compound containing 28.6 wt. per cent filler has dropped to 3.2, implying also considerable reductions in M_c (since theoretically $(1/SR)^{5/3} = \frac{a}{M_c} - 2bM_c^{39}$). Thus, the apparent chain segment between cross-links is shorter than in the unfilled stock (the two were cured to give closely similar degrees of primary valence cross-linkage) and correspondingly the steady-pull modulus is higher. Yet, the internal friction, while also higher, seems to reflect new relaxations from interaction with the filler, and total shock-absorbing power has declined.

Microcrystalline Polymers

The preceding studies at comparatively low frequencies indicated (1) magnitudes of shear rigidity and internal viscosity characterizing rubbers and soft plastics. By familiar shifts of temperature or frequency, they would also apply to polymers known as hard, amorphous plastics at room temperature such as polystyrene and polymethyl methacrylate. (2) Dispersion of μ and μ' with frequency affirm that the intrinsic or fine structure relaxations have times $<10^{-3}$ to 10^{-4} sec, and so refer to chemical units much smaller than the average molecules in the usual technical rubbers and plastics. A way to get at what sizes and habits these units might have will be by investigation of low molecular weight polymer liquids. But, while still in the section on solids, it is recalled that microcrystalline polymers such as polyethylene, polyesters (Terylene), polyamides (nylons), cellulose esters, polyvinylidene chloride, polyacrylonitrile etc., have mechanical properties dominated by their crystalline-amorphous ratios.^{9, 26, 40, 41} The amorphous volumes are clearly those which donate the flexibility, toughness and shock-resistance of these plastics and textile fibers.^{9, 40} An interesting point is, how "viscous" are the disordered chain segments? In an over-all sense, all kinds of dissipation including crystallite friction, analogous to grain friction in metals, scattering of longitudinal waves, and stiffening by low temperatures can occur in these polyphase systems. Thus, effects of chain orientation as well as lateral order (crystallinity) have been detected in dynamics studies.^{26, 41} The intrinsically liquid-like or amorphous components of this behaviour—and the things which will correlate most simply with dipole concentration and other chemical features—are most accessible to study at very high frequencies. For, in these polymer solids, unlike the essentially continuous and homogeneous amorphous ones first discussed, the mechanics reflect small regions having widely divergent properties. Thus, methods developed by H. J. McSkimin of Bell Tele-

phone Laboratories (described in the last issue), have been used to probe for elemental reactions at the upper end of the frequencies presently available. Both longitudinal and shear waves were used. In polyethylene, a wavelength for the shear waves was 0.0074 cm., at $f = 8.55 \times 10^6$ cycles, and in polyhexamethylene adipamide (the usual 6-6 textile nylon), the shear wavelength was 0.0125 cm., for $f = 8.67 \times 10^6$ cycles, all at 25°C.

The important consequence of these experiments so far has been that, despite the small strains involved, the viscosity appears to be a "polymer" viscosity, rather than an inner friction involving just a few liquid-like atoms per unit. Thus, polyethylene of "equilibrium" crystallinity and average molecular weight corresponding to an intrinsic viscosity in xylene of $[\eta] = 0.89$ (at 85°C), was measured over the range from 0 to 50°C. The results from both longitudinal and shear wave measurements

TABLE III

Temp., °C.	f , cycles/sec	Viscosity Poises		
		μ'	$\lambda' + 2\mu'$	λ'
0	8×10^6	15	38	8
0	25×10^6	5	14	4
30	8×10^6	15	34	4
30	25×10^6	5	13	3

are given in Table III. These viscosities are expressed in this case for a Kelvin-Voigt model, of rigidity and viscosity *in parallel*. The rigidities associated with these viscosities are about 3×10^9 dynes/cm², or not far from the value under steady pull of about 1×10^9 .

Now this suggests that the rigid plastic polyethylene retains, even under mechanical impulse of microsecond duration, a shock-absorbing capacity reflected in a shear viscosity of 5-15 poises, and a compressional viscosity of 3-8 poises. The former, μ' , may roughly correspond to the liquid viscosity of a paraffin-like chain of from 50 to 65 c-atoms in length. Thus, the dynamics measurements seem to relate to basic premises of polymer structure. These are that the amorphous regions (whose existence is shown quite independently by x-ray scattering, density, heat-capacity, etc.) indeed take up and dissipate sudden stresses which the microcrystallites, despite their great strength, would be too brittle to sustain.

These results give hope that further probing of the dynamics of liquid-like elements in rigid plastics will eventually lead to precise adjustment

of molecular weight, chemical structure (degree of branching in polyethylene), crystallinity, etc. These quantities, when fitted to a given pattern of μ , λ , μ' and λ' at proper frequencies would yield plastics of optimum serviceability under the multitude of stresses encountered in use.

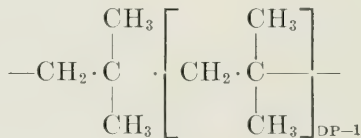
A similar liquid-like structure—even where the (crystalline) rigidity is much higher⁴⁰ and mobile chain segments smaller—apparently occurs in polyamides. Presumably the hydrogen bonding and dipole interactions are very imperfect in the disordered regions,⁴⁰ and there the chain interaction is reminiscent of polyethylene. For instance, in polyhexamethylene adipamide, measurements in the 8 to 30 megacycle range do indicate that the Lamé elastic constant λ is about 5.6×10^{10} dynes/cm², but only about 3×10^{10} for polyethylene. This reflects over-all stiffness dominated by crystallites. Nevertheless, the compressional viscosity, λ' is 17–6 poises (going from 8 to 30 mc) for the polyamide, but only 5–2 poises for polyethylene. Of course, since there is dispersion in both cases, these relative magnitudes might be quite different at some other frequency or temperature (all above are at 25°C). Yet it remains that the nylon, despite its hardness, also has a liquid-like component more viscous than that of polyethylene. Similar relations appear in the shear viscosities, μ' , also determined for these two systems. For the 6–6 polyamide, μ' goes from 19 to 7 poises over the 8 to 30 mc interval while polyethylene changes from 15 to 5. These quantities indicate again, as with the polyethylene, that “polymer liquids” rather than just a few small groups of atoms are the important mechanical elements even at frequencies of 10^7 . Now polystyrene, an amorphous polymer, also has rigidities of about 10^{10} dynes/cm² but the μ' and λ' values at room temperature are far below 5 to 20 poises, and glass-like brittleness (although not so bad as silica glass) results.

So far, then, the two characteristic extremes of polymer mechanics have been discussed: (1) the soft rubbers, whose dynamics at low kilocycle frequencies imply, at ordinary temperatures, predominantly overlapping combinations of relaxation processes whose relaxation elements involve many segments per molecular chain; and (2) the hard, microcrystalline plastics, whose behaviour is predominated by relaxation processes having times of 10^{-6} to 10^{-7} sec because the longer period (slower) displacements have been relaxed out at the temperatures of normal use. (Likewise, interconvertability by temperature¹⁹ between these two extremes is presumed. Also, a certain correspondence between dielectric and dynamic relaxations in these classes is indicated.^{41a}) Next, it is in-

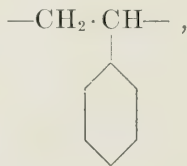
teresting to see what are the simplest structures (particularly in terms of molecular weight) yielding these effects. In other words, what kind of liquid really exhibits "polymer mechanics?" No detailed answer to this can be given below, but results on some polymer liquids of low average molecular weight will indicate that the mechanisms in rubbers and plastics are probably more general than previously supposed.

POLYMER LIQUID MECHANICS

By techniques described in detail elsewhere,^{22, 23} a series of polyisobutylene liquids have been investigated. These polymers were made by ionic catalyzed mass polymerization at reduced, but not very low, temperatures. While no great care to purify the monomer was used, such polymerizations require fair purity to go at all. Seemingly, the resulting liquids do represent a polymer homologous series, although head-to-tail sequence of the monomer units, some single ethyl rather than paired methyl side groups, etc., may differ slightly from the higher molecular weight forms in Butyl rubber and polyisobutylene gum. Whatever are these details, it appears that the polymers represent a linear hydrocarbon chain, with essentially two methyl groups on every other chain atom:

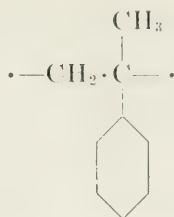


By contrast, polyethylene, with the nominal chain $\text{---CH}_2\text{---CH}_2\text{---CH}_2\text{---CH}_2\text{---}$ and to a lesser extent polystyrene,



have chains in which rotation about the bonds is less sterically hindered. The final section, on isolated polymer chains (in dilute solution), will consider this aspect further. However, some results will be reported below on a low molecular weight poly- α -methyl styrene, which may be considered structurally a cross between the rubber, polyisobutylene, and

the plastic, polystyrene,



Other studies in progress on liquid polybutadiene, polyisoprene, polypropylene, and polypropylene sebacate from which further information about intra-chain stiffness may be derived, will also be noted.

Properties of the polyisobutylenes studied are summarized in Table IV, some additional molecular weights in this range appear as extra points in some of the high-frequency graphs. The molecular weights \bar{M}_η are "intrinsic viscosity" averages^{42, 42a} and, with reasonable estimations of the $\frac{\bar{M}_\eta}{\bar{M}_n}$ ratio, check with cryoscopic number average, \bar{M}_n , values on such materials, which are in turn listed in the table as expressed by melt viscosity relations of Fox and Flory.^{42a} These molecular weights repre-

TABLE IV

Polyisobutylene Polymer	\overline{DP}_η	\bar{M}_η	\bar{M}_n	25°C viscosity	Maxwell μ	Voigt μ'	Maxwell μ'	Freq. Cycles
				Poises		Poises	Poises	
A	10	565	318	0.37	3×10^8	0.6	0.6	14×10^6
A''	30	1660	697	39.6	6.2×10^6	16.5	18.8	2×10^4
					1.7×10^9	7.9	10.0	14×10^6
B	45	2520	1070	216	3×10^9	15.2	24.2	14×10^6
C	56	3350	1720	737	3.6×10^9	20.2	47.9	14×10^6
D	74	4170	2530	1840	4.5×10^9	23.4	78.9	14×10^6
E	147	8240	4850	4600	5.3×10^9	27.2	92.3	14×10^6

sent reasonable averages rather than absolute values for these heterogeneous polymers. The \overline{DP} values are just the number of isobutylene units per average chain. The η values are the steady flow viscosities at low rates of shear—usually determined by a falling ball.

Rigidity and Viscosity Magnitudes

The properties of these liquids ranging from polymer A having only forty times the viscosity of water to E, which begins to approach fluidities of technical polymer melts (polyamides, for instance), were explored

in the kilocycle range with shear waves generated by torsional crystals and in the megacycle region by shear waves with the reflectance method and by longitudinal (ultrasonic) waves from a pulse propagation technique.²³ The results have been expressed in two ways. First, in earlier reports,^{22, 23} a trend corresponding with experiment was given by two Maxwell elements arranged in parallel. This result is too simple compared to the distributions of relaxation times previously proposed for high molecular weight polymers to reproduce *detailed* observation. Nevertheless, perhaps because of the smaller molecules involved, there seems to be clear indication that two *principal* relaxations predominate the mechanical reactions of these liquids over the range of frequencies of present interest, 10^2 to 10^7 cps. For example, for polymer D, these are:

First Relaxation	Second Relaxation
$f_c \sim 4 \times 10^3$ cycles $\mu \sim 4 \times 10^7$ dynes/cm ²	$f_c \sim 5 \times 10^6$ cycles $\mu \sim 6 \times 10^9$ dynes/cm ²

(In accounting for the second main relaxation, a hysteresis component had to be introduced whose significance has been suggested.²²)

Second, specific values of shear rigidity μ (Maxwell) and μ (Voigt), shear viscosity μ' (Maxwell) and μ' (Voigt) as well as the constants for related compressional wave systems, $\lambda + 2\mu$ (elastic) and $\lambda' + 2\mu'$ (viscous) have been calculated for particular frequencies. Unlike in the first way of expression, these latter quantities are all highly frequency dependent. However, they describe conditions at various frequencies of interest, and are thus often worthwhile.

Both ways of looking at the data lead, as implied by the figures above, to the proposal that typical polymer *stiffness* (shear rigidity of $\sim 10^7$ dynes/cm²) is present at $\bar{M}_\eta \sim 1600$, with $\bar{DP}_\eta \sim 30$, or an average chain length of about 60 carbon atoms. This appears when the straining is done in 10^{-3} to 10^{-4} sec. In the 10^{-6} to 10^{-8} sec range, rigidity occurs for even an average chain length of 20 atoms as shown in Table IV.

STRUCTURAL FACTOR IN LIQUID MECHANICS

The main relaxations in the kilocycle range in polyisobutylene liquids seem to lead to quasi-configurational elasticity. This is where the kinetic theory tendency for a most probable separation of chain ends is retarded by viscous interaction of segments between and within the chains. Hence, the middle dashed curves of Fig. 5, showing shear elasticity for some of the polymers of Table IV, decrease exponentially with increasing tem-

perature. While pure kinetic theory elasticity would give a modulus *increasing* linearly with increasing temperature, these systems, like all practical rubbers and plastics, actually grow softer with rising temperature when deformed dynamically. It is striking, nevertheless, that a modulus of $\sim 10^7$ dynes/cm² seems characteristic of the visco-elastic energy storing of these simple polymer structures. As noted below this is 10^3 less than the crystal-like, close-packed, stiffness found for these same molecular frequencies above their second principal relaxation time.

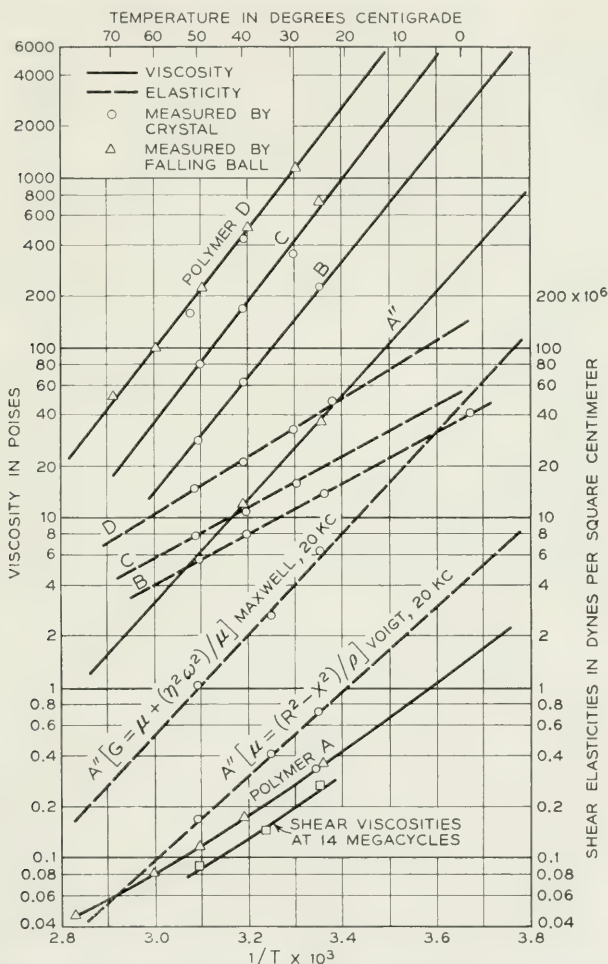


Fig. 5—Shear elasticities and viscosities of polyisobutylene liquids assuming a Maxwell model with relaxation frequencies 10^3 – 10^4 . (Lower dashed curves for frequency-dependent models.)

Thus, it seems that the former, 10^7 , modulus is typical of the structural arrangements in polymers said to be above their second order transition temperatures^{12, 13} while the second, 10^9 , modulus reflects interactions below the freezing-in.

These conclusions obtain regardless of the particular expression of the data. But, for comparison, curves are shown on Fig. 5 for a polyisobutylene A" in which the dynamic modulus μ at 20 kc is computed for both Maxwell and Voigt elements. The two points denoting the steady flow viscosity of polymer A" rank it with respect to the others in the series.

Apparently even very fluid polymer melts, chain molecule plasticizers, and small segments of long molecules must be expected to show appreciable rigidity when stressed rapidly. Referring to the introduction it is reasonable that rough extrusions, frozen-in molding stresses and the like are so easily produced. The lines of Fig. 5 are not, of course, implied to be linear over any considerable temperature range. In the region represented, the temperature coefficient for viscous flow is about 16 kcal for the B, C and D liquids (about 12 for A). This agrees roughly with the steady flow values found for very high molecular weight polyisobutylene.^{38, 42a} The temperature coefficient for the rigidity is less, as would be expected, since the whole center of gravity of the chain need not be displaced, but only local segments.

This quasi-configurational elasticity is increased by molecular weight (although kinetic theory elasticity of chain segments in a network is decreased by increasing segment length). The log μ vs density at 25°C plotted in Fig. 6 indicates that the chief influence is the number of chains per cc, since the points for all the molecular weights now lie on a single line. It should be repeated that the elasticity modulus plotted, μ , is again for a roughly frequency-independent or "absolute" model.^{22, 23} The same is true for the three solid lines on Fig. 6, showing μ in the second, or 10^7 cycle, relaxation range. Here effects of detailed liquid structure come out; the three average molecular weights no longer lie so nearly on a single line. This elasticity is presumably from the crystal-like interaction of nearest-neighbor segments. If temperature is adjusted so that densities are the same, it is seen that the *lower* average molecular weight liquid has the *higher* elasticity modulus. This difference is not large, and should not be interpreted as showing an equal segment interaction, for a polymer of lower specific volume (B compared to D). Rather, it emphasizes in this relaxation range, approaching the "glass" behaviour, that the relaxation rate is vastly more temperature dependent than the specific volume change alone, and structural variations in the

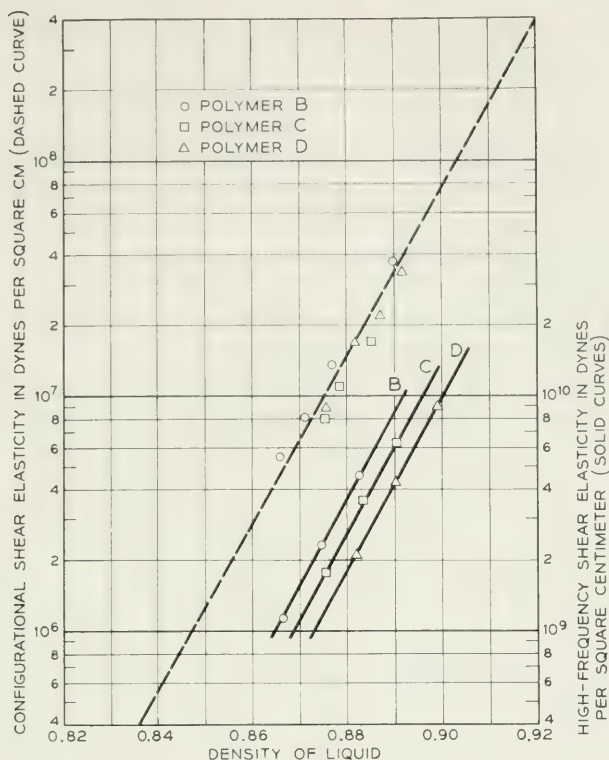


Fig. 6—Principal shear elasticities of polyisobutylenes as related to their densities.

packing of segments in the liquid (coordination number, etc.) become important.

“Soft” and “Hard” Liquid States; Second Order Transitions

A recent noteworthy study⁴³ of volume-temperature and viscosity-temperature changes in polystyrene (with a note on polyisobutylene) brings out many points in common with ideas of polymer liquid structure indicated by the dynamics work.^{22, 23} Particularly, the fact that according to steady state measurements, the “local configurational arrangement of the polymer segments”⁴³ below T_g remains fixed accords with the postulations from dynamics work. That is, above the second main relaxation, it seems to be just the interactions in these fixed arrangements which cause the glassy (or “crystalline”) dynamic modulus of 10^9 to 10^{10} dynes/cm². Further, the point that T_g is *not* an isoviscous state for polymers⁴³ agrees with the dynamics result that macroscopic

viscosity of the polymer has relatively little to do with the actual values of dynamic viscosities. These would be at frequencies where the response of the polymer liquid to the mechanical field is determined only by motions within the local fixed arrangements mentioned above.

Fig. 7 illustrates this, where on one scale the macroscopic viscosity is plotted according to the familiar log-log relation with molecular weight. Two extremes of average molecular weight, \bar{M}_n and \bar{M}_η are used for the liquids, to show that the molecular weight distribution does not alter the general conclusions. (\bar{M}_η is an upper limit weight average figure.) On the other scale, the dynamic viscosity μ' , in this case for a single element *frequency-dependent* Voigt model, shows low values and marked curvature. These betoken the relaxation in which molecular weight, through its effect on free volume and other structural factors, is signif-

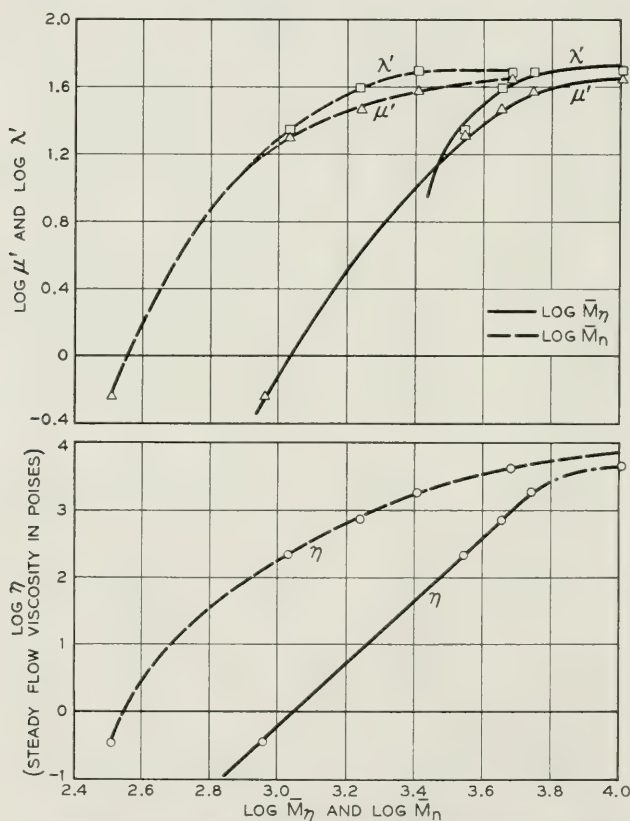


Fig. 7—Comparison of steady flow and dynamic viscosities (at 25°C and 8 mc) of polyisobutylene liquids of different molecular weights.

icant for displacements superficially quite different from those in macroscopic viscosity.

The compressional viscosity, λ' is also plotted, for the same model, in Fig. 7. It is, within experimental error, zero for polymer A', as determined by shear and compressional wave studies at 8 mc frequency.²² This is a rare case, then, where the attenuation of sound waves through a liquid has been quantitatively accounted for by the shear viscosity. But, as soon as the average molecular weight rises to 1000 or so, λ' comes in clearly, and the new mechanism for dissipating compressional or dilatational stresses is developed. As this presumably represents directly free volume or coordination number changes in liquid structure,^{44, 45} its detailed study near T_g ,⁴³ and in connection with brittle points of rubbers, may eventually be especially fruitful.

Another depiction of influence of average molecular weight in these liquids on dynamic viscosities occurs in Fig. 8. Here, the λ'_c curve is

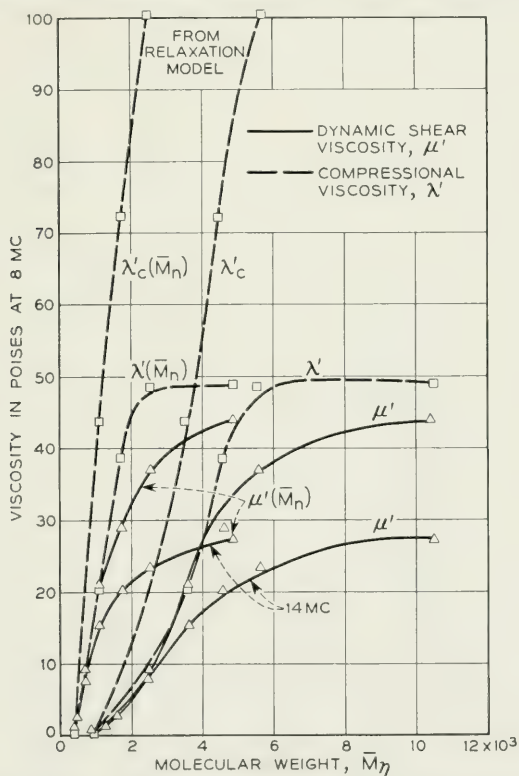


Fig. 8—Dynamic viscosities of polyisobutylene liquids as function of average molecular weight (25°C).

for, again, a crude model attempting to show compressional viscosity over the whole frequency range, while the other viscosities are Voigt expressions at 8 (or 14) mc. Extremes of molecular weight averages are shown.

Comparison of the "soft" or quasi-configurational rigidities, expressed, like the μ of Fig. 5 as relatively frequency independent μ_c , with the "hard" or glassy rigidities is given in Fig. 9. The λ and μ values are for the Voigt model at 8 mc. The graph does not show the bend-over of the "soft", μ_c , curve with molecular weight, but that happens more gradually. The "hard" rigidities λ and μ quite readily show this inflection. As before, the relaxing segments must be <100 chain atoms, according to the behaviour of the molecules at room temperature.

Concerning influence of molecular weight on engineering "brittle points" of such importance in rubber technology, the present studies agree with earlier proposals. Thus, although the T_g or v - T second order transition point always decreases with decreasing molecular weight,⁴³

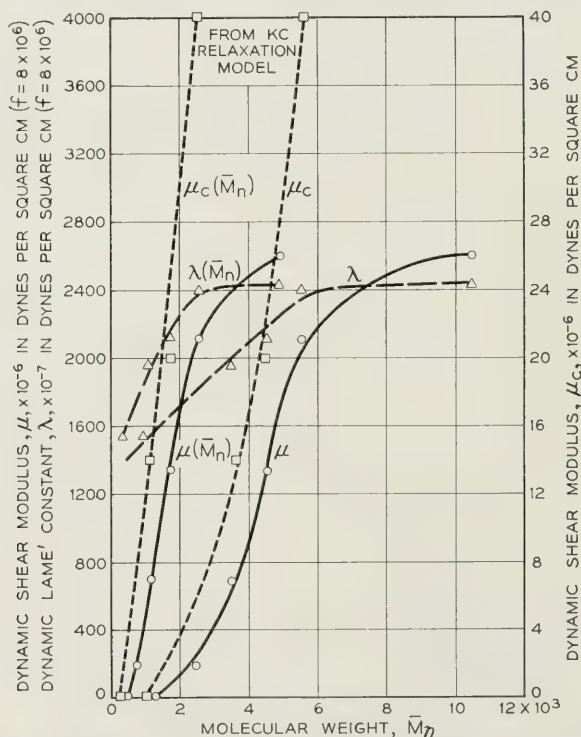


Fig. 9—Dynamic rigidities of polyisobutylene liquids as function of average molecular weight (25°C).

the brittle point tends to increase as the molecules get smaller. This was supposed to be because the ultimate elongation (doubtless due to visco-elastic or quasi-configurational elasticity and not kinetic theory elasticity as sometimes said) declined with chain length, so that the specimen broke at lower and lower strains,¹³ even though it was not really in the glassy state. Now, the results above^{22, 23} demonstrate that the shear modulus at a given temperature does fall off for average molecular weights of polyisobutylene below ~ 5000 . Hence, the mechanical embrittlement of the low molecular weight samples is not because they are stiffer, but because they are weaker.

POLY- α -METHYL STYRENE: A "PLASTIC" LIQUID

Most of the liquid studies have been on polyisobutylene polymers, made "hard" or "soft" by temperature or frequency, but under use

TABLE V

Poly- α -methyl styrene	η Poises	μ' , 14 mc, Poises		μ , 14 mc dynes/cm ²	
		Maxwell	Voigt	Maxwell	Voigt
I	242	111.4	23.6	5.1×10^9	4×10^9
II	4340	502.7	14.3	7.6×10^9	7.4×10^9

conditions, considered rubbery. If one of the methyls in polyisobutylene is replaced by a phenyl, poly- α -methyl styrene, a hard plastic is produced. Low molecular weight polymers of this composition are, however, liquids at room temperature. Hence, it is interesting to compare their reaction to mechanical waves with that of polyisobutylene liquids of similar macroscopic viscosity. Table V lists a few properties at 25°C.

Polymer I has roughly the steady flow η of polyisobutylene B. Also, the Voigt μ' at 14 mc is similar: 15.2 poises compared to the 23.6 poises of the poly- α -methyl styrene. However, the Voigt μ is already 50 times higher for the phenyl substituted chain. (This shows the shift of the *second* relaxation range, of course, where nearest neighbor interactions rule.) Even more striking, the temperature coefficient of a μ'_c , calculated as the second principal shear viscosity in a frequency independent model,²³ as before, is about 24 kcal, compared to about 12 for μ'_c for polyisobutylene in its similar model. Thus, although there was no apparent difference between the mechanical properties of these two polymer liquids in their room temperature state, their dynamics diverged remarkably. This was when they were studied with shear waves whose

frequencies approached the glass-into-rubber relaxation times. Clearly, again, individual interaction of chain-like chemical units and not any micellar or other special aggregation of them, predominates polymer mechanics.

It still remains, however, to separate interactions of the basic units within and between chains. Most likely, the model plastic vs rubber liquids just discussed differ in the high frequency region substantially *only* because of *inter-chain* forces between phenyl vs methyl groups. However, especially in the high-frequency region, questions of *intra-chain* structure, such as the steric hindrance of adjacent pairs of methyl groups in polyisobutylene, restricted rotations about bonds, etc., come in. Obviously where configurational or quasi-configurational displacements are important, as in all cases of elongation >20 per cent (this is certainly an *upper* limit), flexibility of single chains needs to be understood. This is built deeply into chemical structure; plasticizers presumably may change over-all configuration as well as modify interaction, but they are impotent to vary flexibility. Accordingly, problems of rubber, usable in the Arctic, and of wire and cable insulation bendable at low temperatures always come back to whether the polymer chain bonds have free rotation. Some examples of the combinations of effects within and between chains can indeed be shown in several other polymer liquids which are rubber models.

This influence of small changes in chemical structure is compactly illustrated by comparing a few other hydrocarbon polymer liquids with polyisobutylene. Also, rather dilute dipolar groups have been introduced in the linear polyester liquid polypropylene sebacate, whose structure is otherwise like that of hydrocarbons.^{45a} In Table VI, liquids of the given structure with some (unknown) distribution of molecular weights, were studied with shear waves at 77 and 142 kc at a temperature where each had the same steady flow viscosity. The figure chosen was 700 poises, and the temperature range required to adjust to it in the series was 10.9° to 85°C , meaning that the liquids had comparable consistencies at ordinary temperatures.

Despite these similarities under steady stress, the retardation times, τ' , vary three-fold, with the highly substituted hydrocarbon chains, polyisobutylene and polypropylene, the highest. Despite the intermolecular action of the dipoles in polypropylene sebacate, the low polymer has a short retardation time, although its "brittle point" with decreasing temperature is far above that of polybutadiene or even polyisobutylene. Presumably the flexibility around C—O—C bonds rather compensates for increased dipole interaction. Where both low polarity

TABLE VI—Dynamic Properties of Polymer Liquids of Varying Structure.

Polymer	°C	Den- sity, g/cc	η Poisles	Voigt				Maxwell					
				77 kc		142 kc		$\tau' = \frac{\eta}{G}$ sec	77 kc		142 kc		
				η Poisles	G , dynes/cm ²	η Poisles	G , dynes/cm ²		η Poisles	G , dynes/cm ²	η Poisles	G , dynes/cm ²	
Polyisobutyl- ene C	25.40	.8857	700	165.1	3.5×10^7	134.2	6.4×10^7	21×10^{-7}	197	2.2×10^8	173	3×10^8	6×10^{-7}
Polypropylene	85.0	.8248	700	23.6	6.5×10^6	22.0	9.9×10^6	22×10^{-7}	31.3	2.6×10^7	27.5	5×10^7	6×10^{-7}
Polybutadiene	10.9	.8767	700	7.2	5.5×10^6	5.2	6.3×10^6	8×10^{-7}	25.1	7.7×10^6	14.5	1×10^7	15×10^{-7}
Polypropylene Sebacate	41.7	1.0421	700	12.7	9×10^6	8.5	11.6×10^6	7×10^{-7}	40	1.3×10^7	28.6	1.7×10^7	17×10^{-7}

and chain flexibility obtain, as in polybutadiene and the silicones, dynamic properties apparently accord with brittle points in implying small temperature coefficients of relaxation times. In fact, the temperature coefficient for dynamic viscosity of polybutadiene is only about 1.5 kcal, whereas a comparable figure for polyisobutylene and polypropylene is 12 kcal.

The frequency range in which the structural comparisons above were made, resides, as discussed earlier, in the zone of configurational viscoelasticity. That is, over-all shape changes, rather than just nearest neighbor interactions, are predominant even at these comparatively short average chain lengths. Now, other recent studies of polyisobutylene liquids, at 5 to 100 cps frequency, exhibit no rigidity at 25°C and above, although they become non-Newtonian rapidly as temperature is

TABLE VII. *Shear Dynamics of Polyisobutylene A''.*

$$(\bar{M}_\eta = 1660; \eta_s^{25^\circ} = 39.6 \text{ Poises})$$

T, °C	Freq., cps	Voigt		Maxwel	
		μ dynes/cm ²	η Poises	μ dynes/cm ²	η poises
25	266	3.8×10^3	39.2	1.2×10^6	39.3
25	1601	4.8×10^4	38.4	3.1×10^6	39.0
27	25300	5.4×10^5	19.9	1.9×10^7	20.5
27	41390	1.5×10^6	19.9	1.9×10^7	21.5
27	53060	1.5×10^6	18.1	2.6×10^7	19.3

reduced.^{45b} The questions are, where does the configurational elasticity drop out, as frequency is reduced at 25°C; and does it seem reasonable that this dispersion correlates with a shift in frequencies at lower temperatures. Partial answers are given by very recent studies of I. L. Hopkins of Bell Telephone Laboratories. He has equipped the tuning fork vibrator described earlier³³ with two parallel vanes filled in between with a film of polymer liquid. Pure shear properties can be derived from the response of this system. Table VII lists a few typical figures obtained on polyisobutylene polymer A''. These indeed show that the kilocycle relaxation zone (some new data by McSkimin's torsional pulse method are given for it) extends smoothly down to where dynamic and steady stress viscosities are equal. Seemingly there are no new "extra long time" relaxation mechanisms; probably the slow relaxation times sometimes indicated for high molecular weight rubbers are just displacements of this configurational relaxation to long times because of high molecular weight and internal viscosity.

By contrast to the conclusions associated with the data of Table VII,

some observations at low frequencies on isoviscous properties of polyisobutylenes A'' and C indicate nearly *identical* retardation times. Thus A'' at 25°C and C at 61°C have $\eta_s = 39.6$ poises. The η values at 266 and 1600 cps are also about 39 poises, μ at 266 cps is 3800 dynes per cm² for both liquids, and at 1600 cps is from 3.5×10^4 to 4.8×10^4 dynes per cm².

In the final section, mechanical waves have been used to explore dilute polymer solutions, to see how isolated molecules behave, free of interaction with each other.

DILUTE POLYMER SOLUTIONS

Physical Principles in Measurements

Precise information on dynamics of solutions approaching infinite dilution (and thus complete separation of the polymer chains) is desired here. Again these must be shear dynamics; bulk rigidity of ordinary liquids is so high that a few polymer molecules added cause little effect. Dilution is emphasized because even at 1 per cent by volume, high polymer molecule coils frequently interact, especially in "good" solvents. Thus, several workers have detected shear rigidity in polymer solutions, in one case for polymethyl methacrylate of average molecular weight 320,000, at 1 per cent concentration in *o*-dichlorobenzene.⁴⁶ Very low frequencies used (~ 10 cycles) there and in an earlier study⁴⁷ suggest, however, that even here, appreciable entangling of the molecules created a temporary network such as studied by Ferry.^{27, 28} Such was certainly present in the 5 to 18 per cent solutions of cellulose acetate in dioxane measured in one of the earliest observations of shear rigidity in polymer solutions.⁴⁸

Accordingly, since strictly linear, and hence non-interacting, mechanics are sought for the macromolecules in dilute solution, careful evaluation of experiments is essential. Since already it appears that important over-all (quasi-configurational) relaxations occur for, say, polyisobutylene in the kilocycle range, and it is suspected that not *all* of the interactions involved are between chains, the torsional crystal techniques are attractive. The absolute viscosity of these solutions is very low, so the ammonium dihydrogen phosphate crystal whose piezoelectric qualities are appropriate for polymer liquids in the circuits previously noted^{22, 23, 49} is advantageously replaced by quartz.

Detailed electromechanical behaviour of such crystals in the pure liquids cyclohexane and benzene is of first concern. The electric field applied to electrodes on the suspended crystal produces mechanical

torsion generating pure shear waves. These waves may be modified by the environment around the crystal (vacuum, gas, liquid, solid) and react back. Thereby a mechanical resistance, R_M , and a mechanical reactance, X_M , are imposed on the electrical properties of the crystal element in the circuit. This connection comes out as:

$$\Delta R_E = K_1 R_M$$

$$\Delta f = -K_2 X_M,$$

where ΔR_E is the increase in measured electrical resistance of the crystal element in the medium compared to in vacuum (or practically in *dry* air or nitrogen). The decrease in resonant frequency of the crystal element under these conditions is Δf . Thus K_1 and K_2 are electromechanical constants, which fundamentally may be calculated from the dimensions and piezoelectric constants of the crystals.⁴⁹ Now, in simple, Newtonian liquids,

$$R_M = X_M = \sqrt{\pi f \eta \rho}$$

Thus, by carefully measuring ΔR_E (or Δf) on a liquid of accurately known density ρ and viscosity η , at a given frequency f , and a given temperature, the constants K_1 and K_2 may be evaluated without assumptions and approximations of deriving them. Their constancy will then reflect the electromechanical stability of the system. Their behaviour under various conditions will be illustrated below. One further point is that when a liquid or solution does exhibit shear rigidity, or, in other words, if the single large molecules in a dilute solution are able to store energy, then $R_M > X_M$. Hence, in this case, the observed quantities ΔR_E , and especially Δf require particular precision.

In this regard, typical magnitudes of change of f_R between *dry* air and pure cyclohexane, at various temperatures, appear in Fig. 10. Questions often arise as to the arbitrariness of suspension of the radiating crystal, by the fine supporting and lead wires. The effects with the plain wires, in the solid curves of Fig. 10, are somewhat, but not radically, changed when a metal bead is put on, heavily loading vibrations in the wires, as shown by the dashed curves. In Fig. 11, a somewhat larger influence of the loaded support wires is shown for the R_E values, but both curves, by their smoothness and shape over a temperature range where the thermal expansion and other elastic constants of the metal support wires are quite different from those of the quartz crystal, affirm reliability of mounting and electromechanical coupling.

Fig. 10 shows, even for an 80-ke crystal, that Δf for an organic liquid

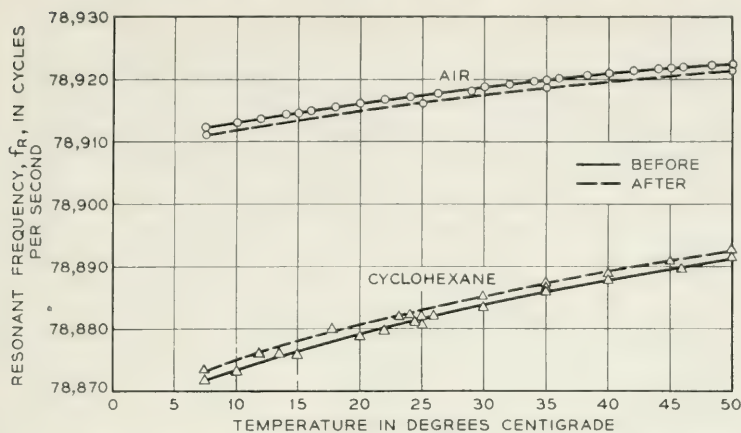


Fig. 10—Temperature variation of resonant frequency before and after adding weights to mounting wires.

(or dilute polymer solution) is bothersomely small. An excellent oscillator at 20 kc can hardly be expected to drift less than ± 2 cycles, but at 20 kc the Δf like that between the sets of curves on Fig. 10 might be only 10 cycles, so 20 to 35 per cent error could come in. Hence, a different scheme for measurement of f_r than that in earlier systems^{23, 49} was evolved. The tenth harmonic of the (say 80 kc) resonant frequency was beat against the 79th harmonic of a controlled standard 10-kc frequency. An interpolation oscillator accurately readable to 1 cycle then supplies the many hundred (roughly 1000) difference between these two high

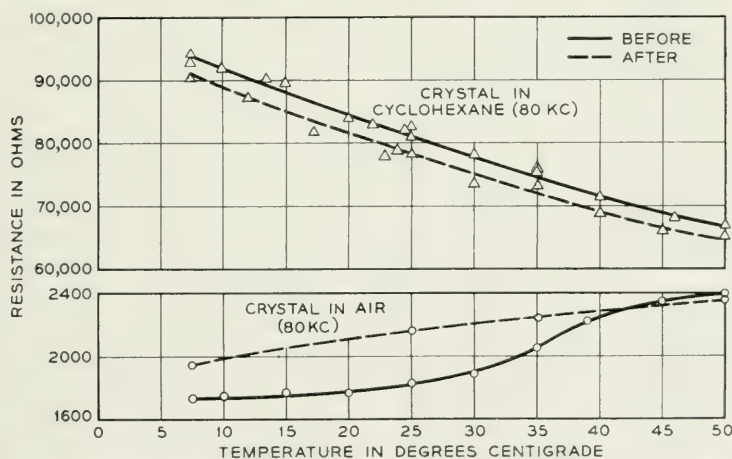


Fig. 11—Temperature variation of resistance at resonance before and after adding weights to mounting wires.

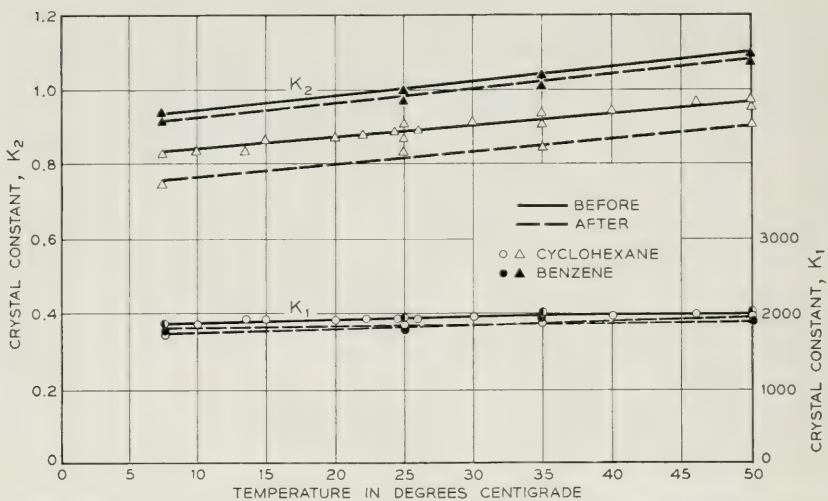


Fig. 12—Temperature variation of crystal constants K_1 and K_2 at 80 kc before and after adding weights to mounting wires.

harmonics. In this way, and in about 30 sec a balance can be conveniently achieved and the required ten-fold gain in accuracy attained.

By these means, and with best literature values of viscosity and density (which were checked in the laboratory at several temperatures) for purified solvents, curves for K_1 and K_2 were obtained as exhibited in Fig. 12 for 80 kc. Behaviour of K_1 at different frequencies over a tem-

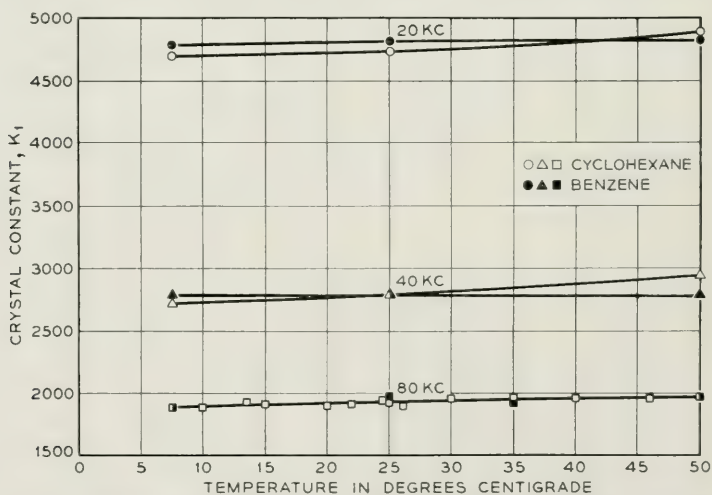


Fig. 13—Temperature variation of crystal constant K_1 over a frequency range with benzene and cyclohexane as standard fluids.

perature range is shown in Fig. 13, and of K_2 , in Fig. 14. Fig. 14 brings out the significant point that in the present arrangement, where the oscillating crystal is immersed in the liquid studied, the *dielectric* properties of the liquid are important. Apparently the dielectric losses even of these purified hydrocarbons are different enough so that K_2 at 80 kc is quite different for benzene and cyclohexane. (Dielectric studies have previously indicated difficulty in preparing benzene having theoretically expected loss.) It is also possible that slight differences in wetting the crystal cause K_2 to vary with the liquid used.

The K_1 and K_2 values determined for all the various conditions above were then used under these conditions for measurements on the polymer

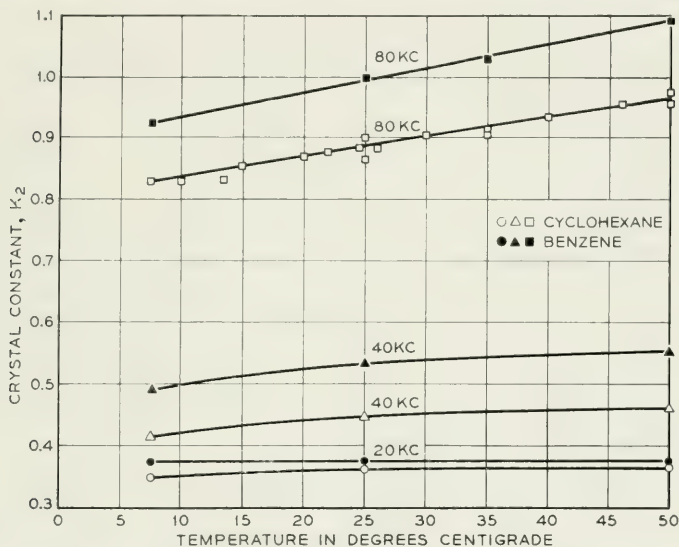


Fig. 14—Temperature variation of crystal constant K_2 over a frequency range with benzene and cyclohexane as standard fluids.

solutions in the kilocycle range. In the megacycle range, the balanced shear wave reflectance technique²³ gave satisfactory results over certain concentration zones which could fairly well be extrapolated to high dilutions. Thus, over the whole spectrum, there seems to be no doubt about the reality of the effects described below. That is, their magnitude far exceeds experimental uncertainty, as demonstrated in this section.

POLYISOBUTYLENE SOLUTIONS; DYNAMICS OF SEPARATE CHAINS

Solutions of polyisobutylene of $\bar{M}_\eta = 1.2 \times 10^6$ from about 0.1 to 1.0 wt. per cent concentration in cyclohexane yield R_M and X_M curves as shown in Fig. 15. The points coincide for the pure solvent, as they

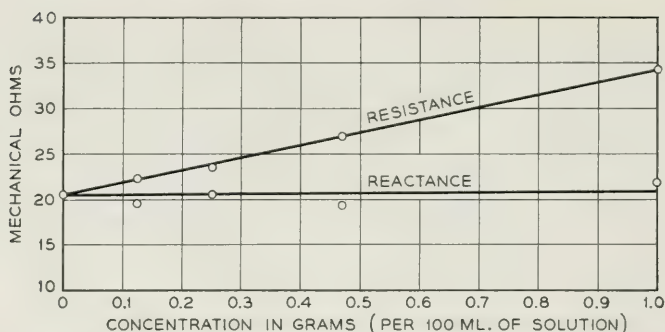


Fig. 15—Electromechanical interaction of solutions of polyisobutylene ($\bar{M}_\eta = 1.18 \times 10^6$) in cyclohexane with crystal vibrating torsionally at 20 kc.

should for a liquid having only viscosity. But, apparently as soon as any polymer chains are added, the curves diverge. A stiffness coming from separate chain molecules is being displayed.⁵⁰ Qualitatively, theoretical expectations of Kuhn^{51, 52} and others seem justified, at least that there is a relaxation mechanism for isolated chains.

The usual question of how best to express the dynamical results arises. The procedure of earlier sections for polymer solids and liquids will be followed. In general, a frequency dependent modified *Maxwell* element as sketched on Fig. 16 will be used. However, a frequency-independent analysis has also been carried out for one sample system, and, from this, basic mechanical constants of single "average" molecules are obtained, if it is reasonable to relate the mechanical models for the liquid continuum to the discrete chains dissolved in it.

Fig. 17 shows typical results from the simple scheme of Fig. 16, where the pure solvent viscosity, η_A , has been considered to be in parallel with a Maxwell element. The total shear rigidity of the solution (at a given concentration) is represented by μ_B . The viscosity of the polymer molecule coils in solution with the solvent streaming through them is

$$\mu_B = \frac{(R^2 - X^2) \omega \eta_B}{\omega \rho \eta_S - 2RX}$$

$$\eta_A + \eta_B = \eta_S$$

$$\eta_A = \frac{2RX}{\omega \rho} - \frac{(R^2 - X^2)^2}{\omega \rho} \cdot \frac{1}{\omega \rho \eta_S - 2RX}$$

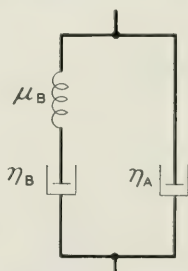


Fig. 16—Relations for calculation of shear stiffness and viscosity of dilute polymer solutions.

taken to be η_B . Thus, the steady flow viscosity, $\eta_s = \eta_A + \eta_B$. Also,

$$\frac{\eta_A + \eta_B}{\eta_A} = \eta_r \quad \text{or} \quad \frac{\eta_B}{\eta_A} = \eta_{rp}$$

under steady flow, or, alternatively, approximately a "dynamic intrinsic viscosity"

$$\left[\frac{\eta_B}{\eta_A} \cdot \frac{1}{c} \right]_{c \rightarrow 0}$$

can be written for any given frequency.

The curves in Fig. 17 are frequency dependent, however, although it turns out that η_B is only slightly so. Nevertheless, the considerable rise of η_A above the pure solvent viscosity, as the concentration is increased, indicates other mechanisms are being lumped into η_A . As usual, some

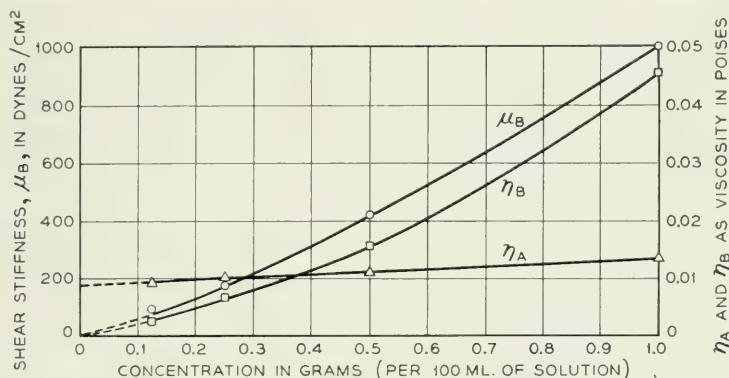


Fig. 17—Rigidity and viscosities of polyisobutylene ($\bar{M}_\eta = 1.18 \times 10^6$) in cyclohexane, at 25°C and 20 kc.

extensive distribution of relaxation times is probably responsible. However, from the chemical point of view, it is best to see if some principal mechanisms related to known structures can be identified. If so, they could be associated with new ideas about the details of polymer intrinsic viscosities, as well as the form of isolated molecules.^{52, 53, 54, 55}

First, the frequency dependence of the μ_B of the model of Fig. 16 is as shown on Fig. 18. Striking regions of dispersion appear, although more points are needed to define the 10^5 cycle zone. Actually, many sets of data have been obtained in the 10^4 cycle zone. Recently, an immersed quartz tuning fork has given the approximate value shown for 2300 cycles. The experiments of Fig. 18 were on a polyisobutylene having $\bar{M}_\eta = 3.9 \times 10^6$, dissolved in cyclohexane. Values of η_A and η_B were, of course, also obtained. The results were then analyzed for a system of

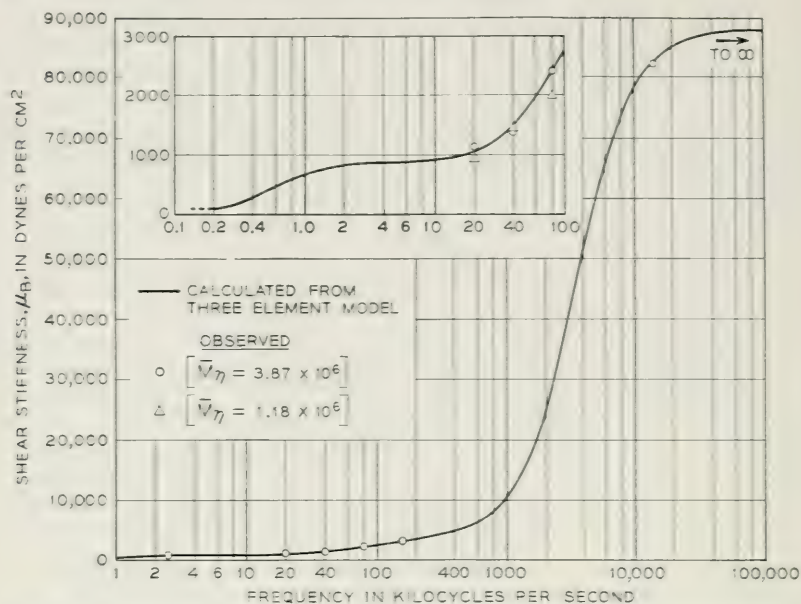


Fig. 18—Frequency dependence of shear stiffness of 1 per cent solutions of polyisobutylene in cyclohexane at 25°C.

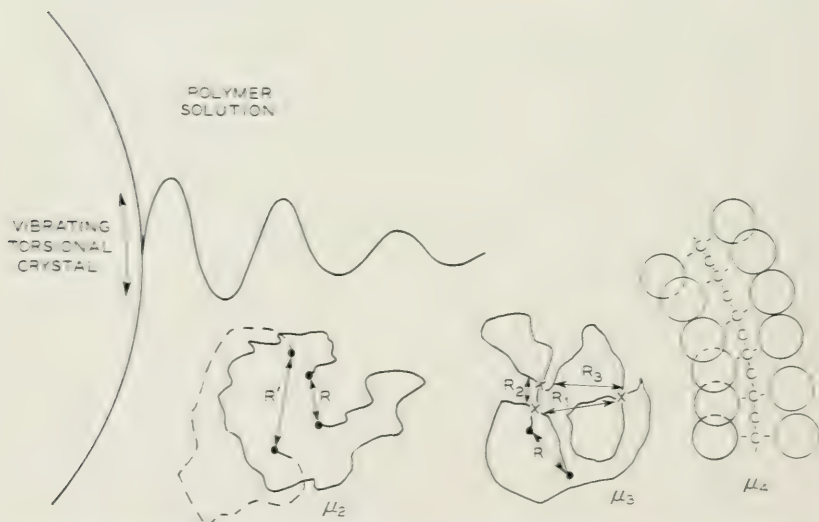


Fig. 19—Schematic diagram of possible sources of rigidity of single chain molecules in solution.

three Maxwell elements in parallel with again, as in Fig. 16, a solvent viscosity, this time called η_1 (truly absolute solvent viscosity) in parallel with them. The μ_B curve of Fig. 18, running through the observed points could then be calculated, by some special trial methods, with which Messrs. H. T. O'Neil and O. J. Zobel of Bell Telephone Laboratories kindly helped.

This analysis, identifying three principal relaxation regions for the motions of polyisobutylene chains in cyclohexane, gave for a 1 per cent solution (taken as linear part of concentration curve and hence equivalent to high dilution).

$$\text{Principal rigidities } \mu_2 = 890 \text{ dynes/cm}^2$$

$$\mu_3 = 3,190 \text{ dynes/cm}^2$$

$$\mu_4 = 84,000 \text{ dynes/cm}^2$$

$$\text{Principal viscosities } \eta_1 = 0.0082 \text{ poise (pure cyclohexane)}$$

$$\eta_2 = 0.255 \text{ poise}$$

$$\eta_3 = 0.006 \text{ poise}$$

$$\eta_4 = 0.004 \text{ poise}$$

$$\text{Principal relaxation frequencies } f_2 = 550 \text{ cycles}$$

$$f_3 = 8.45 \times 10^4 \text{ cycles}$$

$$f_4 = 3.52 \times 10^6 \text{ cycles}$$

Tentatively, these mechanisms may be schematically described as on Fig. 19. Here, the polymer coil, subjected to shear waves in dilute solution, exhibits rigidities μ_2 , μ_3 and μ_4 , all shown on different scales. μ_2 is the configurational elasticity because of actual changes in root mean square separation of chain ends, as from R to R' . It is retarded by viscous drag through the solvent, η_2 , which is presumably the main source of characteristically high η_r of chain polymer solutions. The relaxation frequency for this mechanism is low—a few hundred cycles. It may come in significantly in work on more concentrated solutions at low frequencies,^{28, 45, 48, 57} where chain entanglement is nevertheless the *dominant* factor.

μ_3 is when segments of the same chain in the molecular coil temporarily entangle with each other. Striking evidence has recently been given by Fox and Flory⁵⁸ that because of mutual interference, the theoretical random flight configuration of a chain gives very much too small a

molecular coil volume, V_e . This suggests that thermal agitation tending toward a smaller V_e , and excluded volume or repulsions forcing a large one, will cause collisions or entanglements which might last long enough to give a *van der Waals* cross-bond as denoted by crosses on the μ_3 sketch. (The actual forces in these would somewhat resemble those between different molecules in the concentrated solutions of Ferry.^{28, 37}) This mechanism has the reasonable (based on Ferry's and others' work) relaxation frequency of 8.45×10^4 . A small viscosity, η_3 , may comprise friction of slippage at the entanglement points, with both the polymer and associated solvent molecules.

In Fig. 19, μ_4 is a relatively high stiffness presumed to be some average hindrance to rotation of one segment with respect to another. In the sketch, close-packed spheres representing methyl groups in polyisobutylene are portrayed. Their force fields overlap more in some places than others, in the meandering of the chain to form the molecular coil (of course, some tail-to-tail structures may be important here; they have all been shown head-to-tail in the sketch). Thus, this total internal steric restraint on chain flexibility, with a relaxation frequency of 3.5×10^6 , contributes greatly to the large dispersion of rigidity in the megacycle range noted in Fig. 18. The related viscosity, η_4 , is again low.

There is no doubt a considerable distribution of relaxation characteristics associated with each and all of these mechanisms.

Physical Properties Per Molecule

Since the viscosities and rigidities in the dilute solutions indeed seem to be additive with the number of molecules present, values of these properties, for the hypothetical mechanisms, can be expressed per average chain. Of course, the measured quantities are expressed as constants per cc of solution, but it may be useful to think of in terms of one average chain in each cc. Then, the shear deformation of this chain could be denoted by a force constant. The associated viscosities remain, however, dependent on solvent surroundings. Thus, for the polyisobutylene of $\bar{M}_\eta = 3.9 \times 10^6$, in cyclohexane solution, at 25°C the molecular quantities are:

$$\begin{array}{ll} [f_2] = 17 \times 10^{-13} \text{ dyne cm} & [\eta_2] = 1.6 \times 10^{-16} \text{ poise} \\ [f_3] = 6 \times 10^{-12} \text{ dyne cm} & [\eta_3] = 3.9 \times 10^{-18} \text{ poise} \\ [f_4] = 16 \times 10^{-11} \text{ dyne cm} & [\eta_4] = 2.4 \times 10^{-18} \text{ poise} \end{array}$$

In the section on polymer liquids, the high-frequency modulus μ was attributed to a nearest-neighbor glass or crystal-like interaction (since the actual values were indeed typical of the hardest organic solids).

However, in polyisobutylene (and to some degree in poly- α -methyl styrene), it is especially difficult to distinguish *inter*-chain from *intra*-chain crowding of methyl groups. Thus, while average center-to-center separation of methyls is $\sim 4 \text{ \AA}$ in adjacent chains,⁴⁰ it is $< 2.5 \text{ \AA}$ within chains, in polyisobutylene. This crowding is apparently strong; the observed ΔH_{pzn} is only 12.8 kcal per mole instead of the 19.2 expected.^{55, 56} The energy of steric hindrance thus amounts to almost half of the actual heat of polymerization. It is reasonable that a large part of the hardness of a mass of polyisobutylene chains, such as in the liquids, should therefore reflect the same mechanism as that for μ_4 (Fig. 19) in the dilute solutions. A rough check on this can be made. A polyisobutylene having considerably lower molecular weight than 3.9×10^6 and thus intermediate between the "liquid" and "solid" ranges, had a Maxwell shear modulus in the megacycle region (14 mc) of $\mu = 5.3 \times 10^9$, at 25°C . The number of molecules/cc, with individual $[f_4]$ given above, necessary to give the observed density of this polymer was multiplied by $[f_4]$, giving $\mu = 2.8 \times 10^9 \text{ dynes/cm}^2$. Accordingly, about half of the observed high frequency rigidity of polyisobutylene, at 25°C , may be calculated from a "molecular constant" embodying intra-chain stiffness.

Much more refined and detailed treatments are required to generalize these "molecular constants" which are after all, as shown below, dependent on using a thermodynamically "inert" solvent. However, much as structurally significant dipole moments can be derived from measurements in dilute solutions, it seems hopeful that macromolecular mechanics can be so elucidated. Also additional structures, such as polypropylene and polydimethyl siloxane compared to polyisobutylene, are currently being studied.

Temperature Variation

Some further behaviour at different temperatures and solubilities of separate chains in dilute solution may now be considered against this background of possible mechanisms. Practically, these studies will bear on processing and properties, lacquers, paints, and casting solutions of polymers, as well as on the other qualities outlined in the introduction. Results may be conveniently discussed in terms of the modified Maxwell single element, with factors η_A , η_B , and μ_B (Fig. 16). Mostly, the kilocycle range, reflecting molecular coil changes, will be of interest. For comparison, it may be noted that at 20 kc, the polyisobutylene whose $\mu_B = 1061 \text{ dynes/cm}^2$ in 1 per cent solution in cyclohexane receives 889 dynes/cm² of this from μ_2 , the retarded configurational mechanism; 169

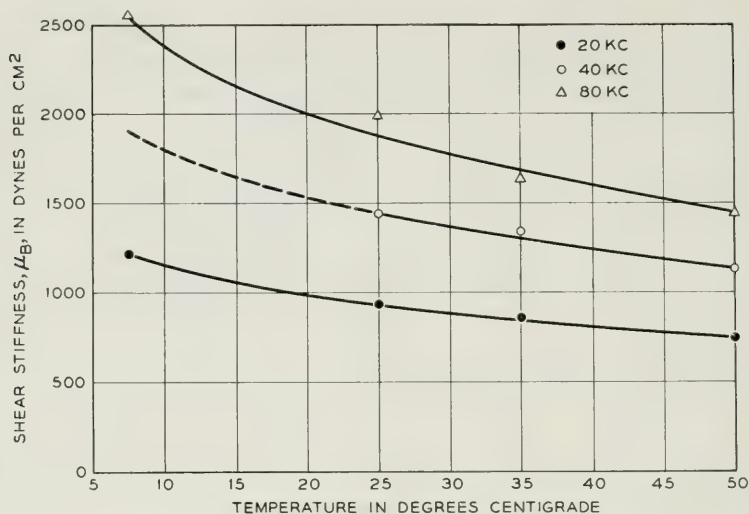


Fig. 20—Temperature variation of rigidity of 1 per cent solution of polyisobutylene ($\bar{M}_n = 1.18 \times 10^6$) in cyclohexane.

dynes/cm² from the entangled segments stiffness, μ_3 ; and only 3 dynes/cm² from the intra-chain stiffness, μ_4 . Thus, chain configuration mechanics, including the associated viscosities, can be well enough thought of in the following paragraphs, in terms of μ_B , η_B and η_A .

The exponential decrease of μ with temperature familiar for polymer solids and liquids is much suppressed in the μ_B vs T curve of Fig. 20. While the μ_4 , internal rotation, mechanism for single polyisobutylene molecules probably has a considerable activation energy, that for the μ_2 , configurational, rigidity should be very small. Then, without retardation, the intrinsic chain modulus would rise with rising temperature. These influences seem to combine to give the modest decline of μ_B appearing in Fig. 20. If these rigidities are plotted against $1/T$, the temperature coefficient is 2.3 kcal. This is much less than the familiar values for the stiffening of rubbery solids, and emphasizes that *inter-chain* action reigns then.

Solvent Variation

Effects of solvents of different (mostly positive) heats of mixing on state of polyisobutylene molecules in solution have been nicely established by Fox and Flory.⁵⁸ Especially, this work has clarified principal factors in the intrinsic viscosity expression

$$[\eta] = \frac{V_e}{M} \cdot \frac{\varphi}{100} = K_0 M^{1/2} \alpha^3 \varphi.$$

Here, V_e = effective volume per molecule (and hence as determined by chain configuration), M = molecular weight, α represents change in linear extent of molecule because of mutual interference of segments^{5h} and φ expresses the hydrodynamics interaction of solvent and molecular coil (including varying degrees of "straining through" the coil).^{61, 62, 63, 64} Interpretation of the mechanical properties of chains in dilute solution, with reference to the rough concepts of Fig. 16, arouses particular interest in the factor α^3 . For a high molecular weight polyisobutylene, intrinsic viscosity theory⁵⁸ indicated that α^3 the ratio for volume of actual coil divided by volume for ideal random flight coil was 3.81 in cyclohexane but only 1.42 in benzene, both at 30°C. This striking alteration in equilibrium chain configuration, a variable which is not readily introduced into polymer liquids or solids, appears in the inherent viscosity vs c curves in cyclohexane, Fig. 21, and benzene, Fig. 22. The large difference in $[\eta]$ at 25°C, 6.00 in cyclohexane vs ~ 1.5 in benzene, indeed emphasizes the different solvent powers.⁶⁵ Likewise, the large increase of $[\eta]$ with temperature in Fig. 22 accents the poor solvent qualities of benzene.⁶⁶ Too, empirically, polymer molecules which are either tight coils or are actually chemically cross-linked to form microgel molecules characteristically show *positive* slopes of inherent viscosity vs c plots.⁶⁷ Accordingly, all this evidence for large changes in the conformation of chain molecules in "good" vs "poor" solvents should show up in dynamics of dilute solutions. Also, technically, quite different physical properties are found for polymer-plasticizer compounds where compatibility is high (good solvent) than where it is low (poor solvent). Here,

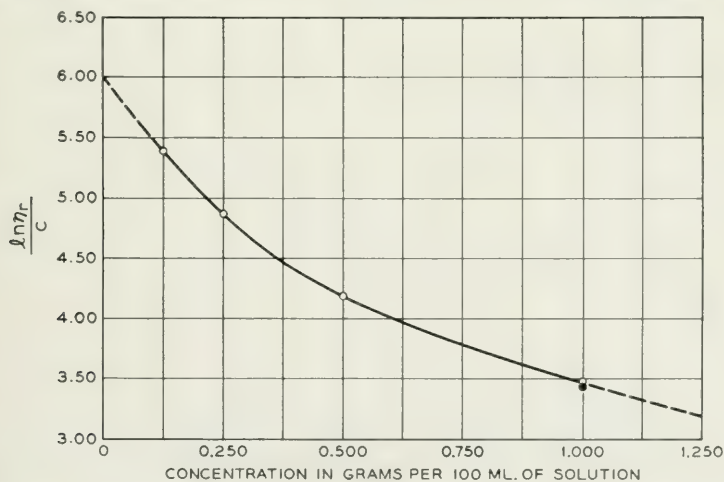


Fig. 21—Inherent viscosity of polyisobutylene ($\bar{M}_\eta = 3.87 \times 10^6$) in cyclohexane at 25°C.

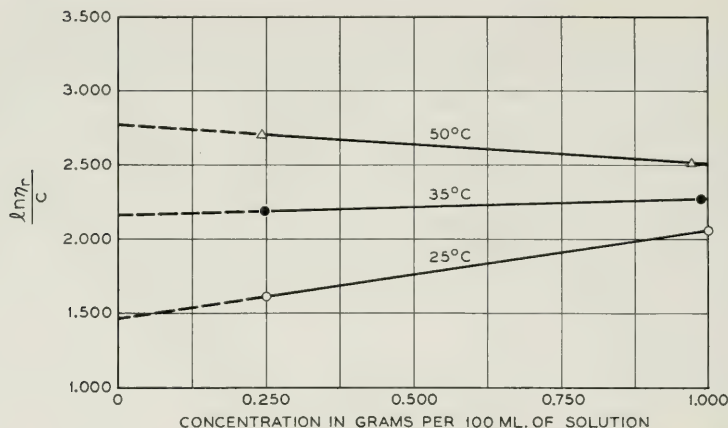


Fig. 22—Inherent viscosity of polyisobutylene ($\bar{M}_\eta = 3.87 \times 10^6$) in benzene, at various temperatures.

more flexible compositions are often produced with low compatibility plasticizers—indeed, sometimes with those on the verge of phase separation than with those with highly favorable heats of solution.⁶⁸ This would mean that the bad solvents would compress the chains so that they would be more easily strained than if they were in a “free chain” or even extended configuration. If single chain, visco-elastic stiffnesses are acting this way, the dynamic μ_B would then actually decline as heat of mixing become more positive.

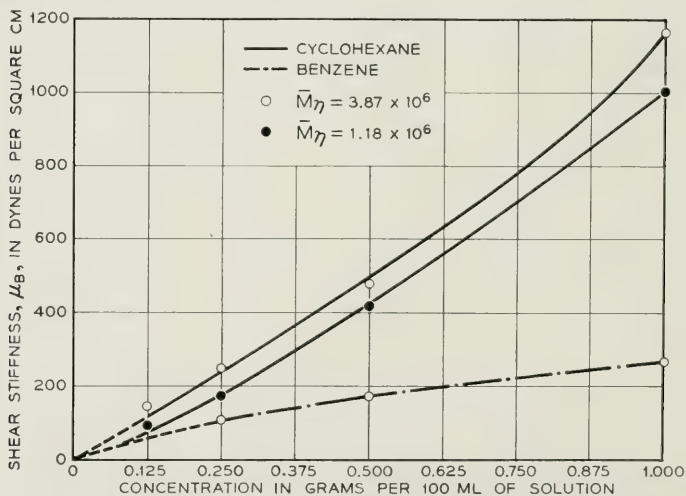


Fig. 23—Rigidity of polyisobutylenes in cyclohexane and benzene at 25°C and 20 kc.

This seems to take place, as indicated by the lower compared to the upper curve on Fig. 23. Here, the μ_B of the usual modified Maxwell model, at 20 kc, is plotted against c for the polyisobutylene of $\bar{M}_\eta = 3.9 \times 10^6$. Also, the middle curve shows μ_B for a polymer of about a third of this molecular weight; while there is a small reduction in μ_B with \bar{M}_η in this range, it is much less than the reduction caused by tightening up the polymer coil.

The μ_B values per average molecule, $[f_B]$, fall from 18×10^{-13} dyne cm in cyclohexane to 7×10^{-13} in benzene. (Of course, $[f_B]$ for the intermediate molecular weight polymer in cyclohexane is only 5×10^{-13} because so many more molecules are present in solution.)

The temperature dependence of μ_B also becomes nearly zero at least over the narrow range from 25 to 50°C, in benzene compared to cyclohexane. This seems to accord with the indications previously, from Fig. 20, for a lower molecular weight polymer, that different mechanisms are competing. These may be the configurational, with $\mu \propto T$, and relaxation, with μ varying in some complicated way with T . Thus $[\eta]$ increases markedly with T , and presumably denotes an expanding molecular coil tending toward the "normal" configuration in cyclohexane. At the same time, the relaxation processes with rising temperature tend to cause the decrease in μ_B typical of the upper, solid, curves on Fig. 24. In engineering use, often times poorly compatible plasticizers give compounds which stiffen more gradually with temperature than do "solvent" plasticized ones.

For similar reasons, the dynamic molecular coil viscosity, η_B , ought to vary less with temperature in thermodynamically poor than in good solvents. This is indeed seen in Fig. 25. On the other hand, η_A for the modified Maxwell element has been described as the solvent viscosity with segment hindrance and restricted rotation terms from the polymer molecules lumped in with it. These latter terms are presumably little affected by over-all configuration (μ_2 term; the μ_3 mechanism will be somewhat affected, but not the μ_4 , on Fig. 19). Thus, η_A should have comparable temperature dependence in both good and bad solvents, as seems to be indicated by Fig. 26.

Microgel Molecule Solutions

The statistical coil of linear polymer molecules may be replaced by a chemically fixed, cross-linked network in microgel molecules.⁶⁷ These may be made completely rigid, like Einstein spheres, or highly swellable. The latter are hybrids between rigid spheres and coiled chains. In

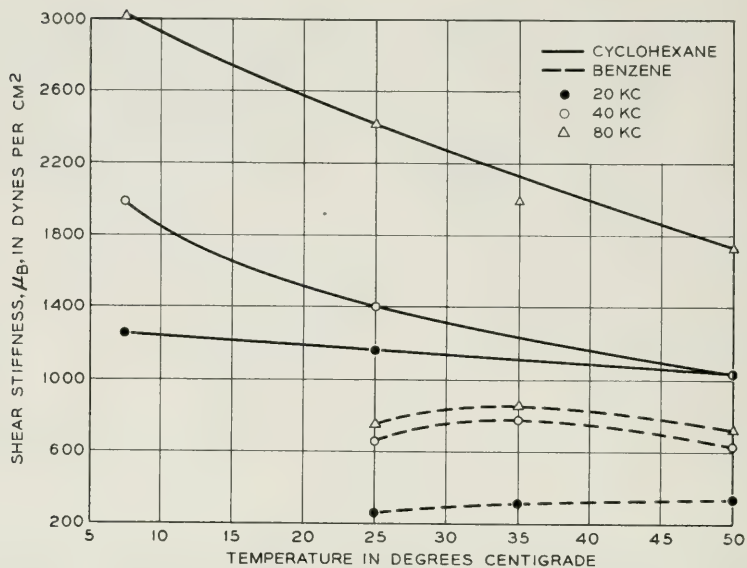


Fig. 24—Temperature variation of rigidity of 1 per cent solution of polyisobutylene ($\bar{M}_\eta = 3.87 \times 10^6$) in cyclohexane and benzene.

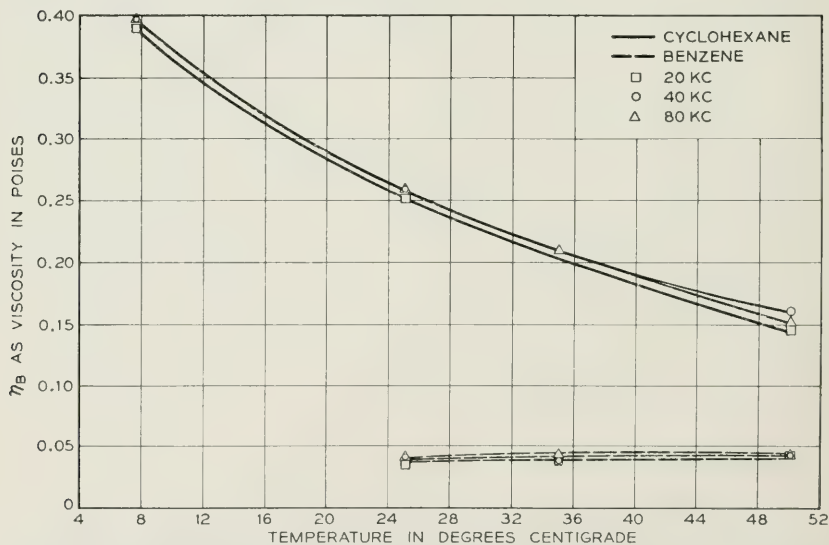


Fig. 25—Temperature variation of η_B for 1 per cent solution of polyisobutylene ($\bar{M}_\eta = 3.87 \times 10^6$) in cyclohexane and benzene.

synthetic rubber, they confer unique flow properties, causing the excellent processibility of GR-S 60. However, dynamic tenacity, such as in flex crack growth, is degraded by their presence. Now presumably the excellent extrusion qualities of synthetic rubber composed of from 60 to 80 per cent microgel molecules are because of their individual shear stiffness. Thus, if a wire coating, for instance, is extruded at high rates of shear, chain molecules are deformed, and store energy just as discussed in the earlier sections on liquids. After emerging from the extrusion die, they relax, and cause the gross retraction, shrinkage and roughness shown in the wire insulation of the upper photograph of Fig. 27. A polymer with about 70 per cent microgel molecules gives the smooth covering shown in the lower specimen of Fig. 27. Here, the shearing stresses of extrusion seem insufficient to distort the tiny networks of the microgel molecule; in any case, the covering does not roughen or relax. Similar effects have been found for microgel plastics. Nevertheless, unlike gross or macro gelation, the whole melt can flow.

On this basis, dilute solutions of microgel molecules ought to indicate high shear rigidity per molecule. The mechanism μ_3 of Fig. 19, in which now the junction points are not temporary, but are primary valence cross-links, should be predominant. Fig. 28 shows, for a polybutadiene microgel in cyclohexane,⁶⁷ that μ_B has indeed risen, compared to equal

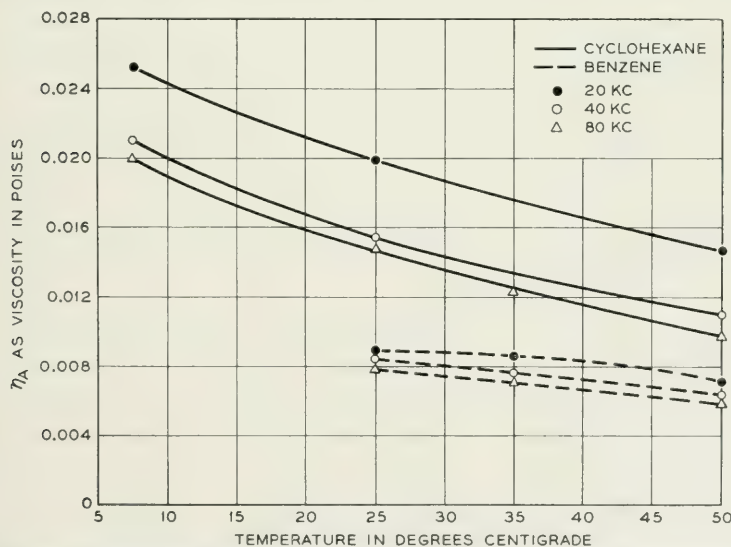


Fig. 26—Temperature variation of η_A for 1 per cent solution of polyisobutylene ($\bar{M}_n = 3.87 \times 10^6$) in cyclohexane and benzene.



Fig. 27—Effect of microgel molecules in synthetic rubber on smoothness of extruded wire insulation. Rough covering is from high-speed extrusion of GR-S without microgel.

weights of chain molecules. Further, accompanying the extremely high average molecular weight of the microgel (18.6×10^6), the $[f_B]$ per average molecule is 42×10^{-12} dyne cm or about twenty-five times that of the polyisobutylene with $\bar{M}_\eta = 3.9 \times 10^6$. Also, the temperature coefficient for μ_B of polybutadiene microgel is low.

Of course, polybutadiene, as chains or as microgel molecule segments, has many double bonds. These will surely influence the μ_4 , or internal rotation mechanism. Further work remains to show just what is their effect in the microgel case. But, it is interesting to compare μ_B values for Hevea rubber chains with those for, say, polyisobutylene, which has only single bonds in the chain.

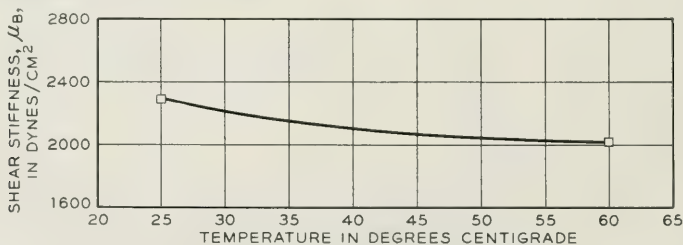


Fig. 28—Rigidity of 0.5 per cent solution in cyclohexane of polybutadiene microgel ($\bar{M}_w = 18.6 \times 10^6$) at 20kc.

Hevea Rubber Solutions

The comparison of equal weight concentrations of natural rubber in cyclohexane with polyisobutylene in cyclohexane is surprising:

Hevea rubber $\bar{M}_\eta = .23 \times 10^6 \mu_B = 1350 \text{ dynes/cm}^2$, 1 per cent solution (corr.).

Polyisobutylene $\bar{M}_\eta = 1.2 \times 10^6 \mu_B = 1000 \text{ dynes/cm}^2$, 1 per cent solution (corr.).

Both results are at 20 kc. The higher value for natural rubber may be because of the double bonds causing stiffening of the chain. On the other hand, maybe easy rotation around single bonds raises the μ_3 part. Certainly the *viscous* retardation *within* natural rubber chains is very low, as noted in the section on solids. However, its interaction with, or configuration, in cyclohexane may be peculiar. The $[f_B]$ per average molecule is, however, low, being 15×10^{-14} dyne cm at 25°C.

Polystyrene Solutions

Much work, on light scattering and other properties, has indicated appreciable intra-chain stiffness for polystyrene,⁶⁹ but still much freedom compared to polyisobutylene.^{69a} However, this work, as well as ΔH_{pzn} of 17 kcal compared to ~ 19 kcal calculated for no steric hindrance, suggests comparatively small restraints on ideal flexibility. This needs to be checked by a frequency analysis of dilute solution mechanics, but polystyrene seems to be a reasonable example of "plastic" behaviour at room temperature because of interaction *between* the chains. (It is recalled that, earlier, α -methyl styrene polymer was cited as plastic model showing both intra- and inter-chain stiffness. Unlike in polystyrene, the intra-chain factor shows up in a ΔH_{pzn} of 9–10 kcal, a third less than that calculated if there were no steric hindrance.) Thus, no evidence of unusual stiffness appears in Fig. 29, when, indeed, the μ_B values are considerably lower, for equal weight concentrations, than those for natural rubber. The highly milled rubber studied had \bar{M}_η very nearly that of $\bar{M}_\eta = 0.234 \times 10^6$ of the polystyrene, so the $[f_B]$ per average polystyrene chain, 4.5×10^{-14} dyne cm is less than a third that of the rubber. No wonder that at high temperatures, where the phenyl group interaction between chains is much reduced, polystyrene makes a good rubber. Also, in Fig. 29 are shown data for a polymer of $\bar{M}_\eta = 1.2 \times 10^6$, made in emulsion and having $[\eta] = 4.350$ in benzene at 25°C.

The polystyrene solutions discussed above were in benzene, a good solvent. Here, the situation is converse to that for polyisobutylene; for polystyrene, cyclohexane is a poor solvent and benzene, good. Hence, if the previous interpretation of reduced single chain *quasi-configurational* (μ_2) stiffness is general for solvents of more endothermic mixing, the "plastic" molecule polystyrene should show it in cyclohexane. This is indeed evident in Fig. 30, showing one of the same polystyrenes of Fig. 29, measured at 20 kc (normalized to 1 per cent concentration). Also, on Fig. 30 are shown the inherent viscosity (practically, the intrinsic viscosity, in this case) and the absolute viscosity of the 1 per cent solution

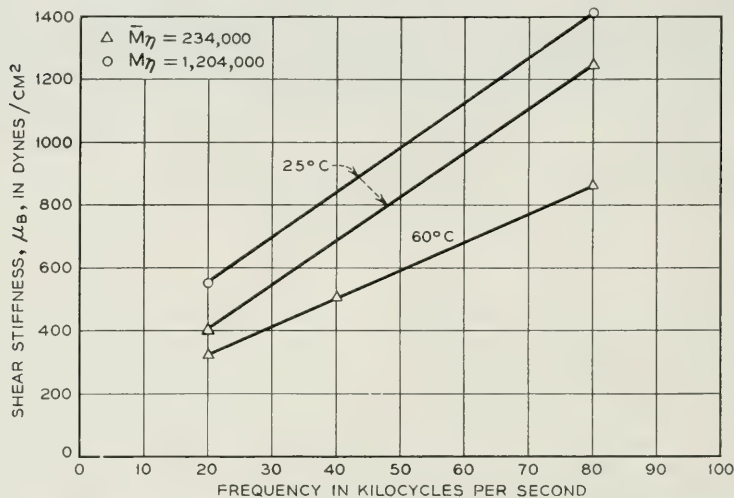


Fig. 29—Change of shear stiffness, μ_B , with frequency, for 1 per cent solutions of polystyrene in benzene.

under steady flow, η_s . These are all plotted against temperature down to phase separation, at about 26° to 27°C.

The marked positive slope of the $\ell n \eta_r / c$ curve denotes the large contraction in molecular coil volume preceding phase separation or insolubility. The absolute viscosity, η_s , however, rises with declining temperature because it is dominated by solvent viscosity, but when the polymer phase comes out, η_s abruptly falls off.

The μ_B values are consistent with this steady flow behaviour, except that the rise of μ_B at the turbidity point seems to be because a layer of swollen polymer-rich phase forms on the torsional crystal surface. This condition is seen in Fig. 30 to coincide nicely with the abrupt changes in steady flow viscosity.

The slight maximum in the μ_B curve at about 35°C may not be real.

It does come near the point of minimum interaction for the whole system. In any case, as discussed before, the average temperature coefficient of μ_B in the poor solvent is very low compared to the good solvent. The values of μ_B are roughly $\frac{2}{3}$ to $\frac{1}{2}$ those in benzene.

GENERAL THEORY OF SINGLE CHAIN MECHANICS; KUHN AND KIRKWOOD

As noted before, much of the present understanding of stress-strain properties of polymer chains, in dilute solutions, liquids or solids, has come from W. Kuhn's long interest in this subject. Many supplementary contributions have been stimulated by Kuhn's work, and new points of view have been introduced by others. For instance, recently new and different proposals have been made about the flow birefringence and non-Newtonian viscosity of solutions of deformable spheres.⁷⁰ These ideas could be tested on suitable microgel solutions.

Recently, moreover, an especially significant general theory of viscoelastic behaviour of polymer in solution has been constructed by Kirkwood.⁷¹ It explicitly considers the hydrodynamic conditions leading to the rigidity now observed for high-frequency shear waves. It formulates definitely the configurational changes of isolated chains in solution when strained in shear. As this theory is advanced to forms where simpler calculations can be made, it may answer many of the questions raised by the new experiments on single chain properties.

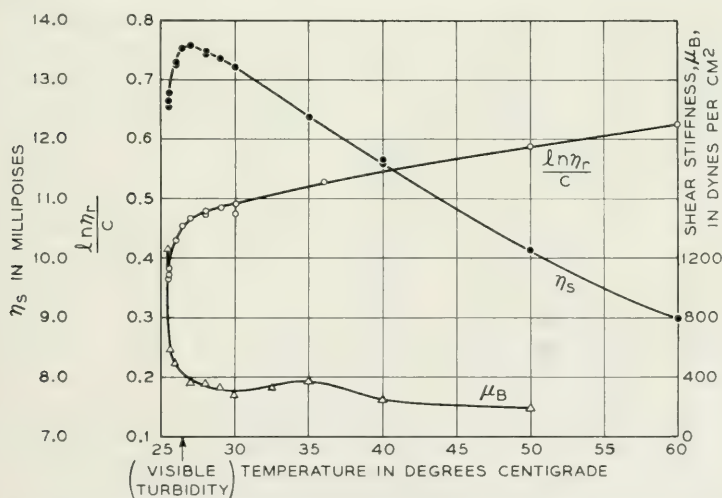


Fig. 30—Temperature dependence of absolute viscosity, η_s , inherent viscosity, $\ln \eta_r/c$, and shear stiffness, μ_B of, 1 per cent solution of polystyrene in cyclohexane down through turbidity point.

CONCLUSIONS

To leave some impression of the elemental chemical structures which move around when wood, rubber, plastics, textiles and finishes are used mechanically—that has been the aim of this study. Polymer viscosities have been found in a variety of “solids”; rigidities have been demonstrated for very fluid “liquids” and solutions. Studies of these solid and liquid extremes have given some chemical reality to the classical spring and dashpot models.

Existence of compressional viscosity has been shown for polymer liquids and solids. It may comprise a new quality for investigation of polymer structure. At present, too little is known of its origin to interpret further the effects of intense ultrasonic irradiation of polymer solutions. Experiments of Schmid and co-workers⁷² early indicated degradation of molecular weight of polystyrene, so irradiated, but whether this is chemical, from local heating in the solvent, or actual physical coupling with the wave field, is still unsettled. However, these workers also considered a compressional stiffness of the polymer molecules in the solutions,⁷³ and showed that if there was coupling, it was not inertial (by dissolving polystyrene in solvents of exactly the same density, no reduction in effect was observed). A point of general interest arises here; impact fractures of plastics presumably actually fracture some primary valence bonds. This is certainly true for many thermoset materials, and probably for chain compounds. Hence, if the detailed mechanism of how compressional waves move and perhaps rupture polymer segments were known, information on the baffling problems of ultimate strength would be gained. The observations above on dependence of λ and λ' on molecular weight and structure provide only the barest start on this but a new goal is in view. Too, basic questions of how rapidly molecules being formed in a polymerization equilibrate in temperature with their surroundings are elucidated by compressional wave propagation constants. For instance, absolute rate measurements on velocity of chain growth cannot be said to be isothermal if they seem to be faster than the thermal relaxation times which the ultrasonic measurements indicate can be $\sim 10^{-6}$ to 10^{-5} sec.

Likewise, more thorough understanding of velocity and dispersion of compressional waves in polymer solutions would clear up anomalies in velocity measurements for a wide variety of polymers,⁷⁴ some of which have been tentatively attributed to chain branching.

ACKNOWLEDGEMENT

Besides the extensive collaboration of W. P. Mason and H. J. McSkimin, we should like to attest to the help of T. G. Kinsley.

BIBLIOGRAPHY

1. Meyer, von Susich and Valko, *Kolloid-Z.*, **59**, p. 208 (1932); Meyer and Ferri, *Helv. chim. Acta*, **18**, p. 570 (1935).
2. Meyer and van der Wyk, *J. Polymer Sci.*, **1**, p. 49 (1946).
3. Kuhn, *Zeit. physik. Chem.*, **B42**, p. 1 (1939).
4. Simha, *J. Appl. Phys.*, **13**, p. 201 (1942).
5. Gilmore and Spencer, *Mod. Plastics*, **27**, p. 143 (Apr. 1950); *Ibid.*, **27**, (Dec. 1950).
6. Mooney, *J. Colloid Sci.*, **2**, p. 69 (1947).
7. Vila, *Ind. Eng. Chem.*, **36**, p. 1113 (1944).
8. Dexter and Dienes, *J. Colloid Sci.*, **5**, p. 228 (1950).
9. Hopkins, Baker and Howard, *J. Appl. Phys.*, **21**, p. 206 (1950).
10. Haward, *Trans. Far. Soc.*, **34**, p. 267 (1943).
11. Morey, *Ind. Eng. Chem.*, **37**, p. 255 (1945).
12. Jenckel, *Zeit. f. Elektrochem.*, **43**, p. 796 (1937).
13. Boyer and Spencer, *J. Appl. Phys.*, **16**, p. 594 (1945).
14. Richards, *J. Chem. Phys.*, **4**, p. 449 (1936).
15. Baker, *India Rubber World*, **110**, p. 543 (1944).
16. Alexandrov and Lazurkin, *Acta Physicochimica USSR*, **12**, p. 647 (1940).
17. Boyer and Spencer, *J. Polymer Sci.*, **2**, p. 157 (1947). An apt survey of much study of plasticizers and mechanical properties.
18. Andrews, Hofman-Bang and Tobolsky, *J. Polymer Sci.*, **3**, p. 669 (1948); Dunell and Tobolsky, *Textile Res. J.*, **19**, p. 63 (1949); Brown and Tobolsky, *J. Polymer Sci.*, **6**, p. 165 (1951).
19. Nolle, *J. Polymer Sci.*, **15**, p. 1 (1950); *J. Appl. Phys.*, **19**, p. 753 (1948).
20. Ivey, Mrowca and Guth, *J. Appl. Phys.*, **20**, p. 486 (1949).
21. Hall, *Phys. Rev.*, **71**, p. 318 (1947).
22. Mason, Baker, McSkimin and Heiss, *Phys. Rev.*, **73**, p. 1074, p. 1873 (1948).
23. Mason, Baker, McSkimin and Heiss, *Ibid.*, **75**, p. 936 (1949).
24. Sack, Motz and Work, *J. Appl. Phys.*, **18**, p. 451 (1947).
25. Lyons and Prettyman, *J. Appl. Phys.*, **19**, p. 473 (1948).
26. Ballou and Smith, *J. Appl. Phys.*, **20**, p. 493 (1949).
27. Ferry, Sawyer and Ashworth, *J. Polymer Sci.*, **2**, p. 593 (1947) for general review.
28. Ferry, *Jour. Res., NBS.*, **41**, p. 53 (1948).
29. Alfrey and Doty, *J. Appl. Phys.*, **16**, p. 700 (1945).
30. Leaderman, *J. Colloid Sci.*, **4**, p. 193 (1949).
31. Kelsey and Dillon, *J. Appl. Phys.*, **15**, p. 352 (1944).
32. Wilson and Smith, *Ind. Eng. Chem.*, **41**, p. 770 (1949).
33. Hopkins, *Trans. A.S.M.E.*, **73**, p. 195 (1951).
34. Rorden and Grieco, *J. Appl. Phys.*, **22**, p. 842 (1951).
35. Gehman, Woodford and Stambaugh, *Ind. Eng. Chem.*, **33**, p. 1032 (1941).
36. Dillon, Prettyman and Hall, *J. Appl. Phys.*, **15**, p. 309 (1944).
37. Marvin, Fitzgerald and Ferry, *J. Appl. Phys.*, **21**, p. 197 (1950).
38. Fox and Flory, *J. Am. Chem. Soc.*, **70**, p. 2384 (1948).
39. Flory, *Ind. Eng. Chem.*, **38**, p. 417 (1946).
40. Baker, Chapter 8 in *High Polymers*, edited by Twiss, Reinhold Publishing Corp., New York, 1945.
41. Work, *Textile Res. Journal*, **19**, p. 381 (1949).
- 41a. Tuckett, *Trans. Far. Soc.*, **40**, p. 448 (1944); Würstlin, *Kolloid-Z.*, **120**, p. 84 (1951).
42. Flory, *J. Am. Chem. Soc.*, **65**, p. 372 (1943).

- 42a. Fox and Flory, *J. Phys. and Colloid Chem.*, **55**, p. 221 (1951).
43. Fox and Flory, *J. Appl. Phys.*, **21**, p. 581 (1950).
44. Debye, *Z. Elektrochem.*, **45**, p. 174 (1939).
45. Galt, *Phys. Rev.*, **73**(2), p. 1460 (1948).
45a. Baker, *Rubber Chem. Tech.*, **18**, p. 632 (1945).
45b. Harper, Markowitz and DeWitt, *Abstracts, XII Int. Congress Chemistry*, p. 274 (1951).
46. Van Wazer and Goldberg, *J. Appl. Phys.*, **18**, p. 207 (1947).
47. Kendall, *Rheol. Bull.*, **12**, p. 26 (1941).
48. W. Philipoff, *Physik. Z.*, **35**, p. 884, p. 900 (1934).
49. Mason, *Trans. A.S.M.E.*, **69**, p. 359 (1947).
50. Baker, Mason and Heiss, *Bull. Am. Phys. Soc.*, **24**, p. 29 (1949).
51. Kuhn and Kuhn, *Helv. chim. Acta*, **29**, p. 609, p. 830 (1946); *J. Colloid Sci.*, **3**, p. 11 (1948).
52. Flory and Fox, *J. Am. Chem. Soc.*, **73**, p. 1904 (1951).
53. Fox, Fox and Flory, *Ibid.*, **73**, p. 1901 (1951).
54. Fox and Flory, *Ibid.*, **73**, p. 1909 (1951).
55. Fox and Flory, *Ibid.*, **73**, p. 1915 (1951).
56. Kuhn and Grün, *J. Polymer Sci.*, **1**, p. 183 (1946).
57. Carver and Van Wazer, *J. Phys. and Colloid Chem.*, **51**, p. 751 (1947).
58. Fox and Flory, *Ibid.*, **53**, p. 197 (1949).
59. Evans and Tyrrell, *J. Polymer Sci.*, **2**, p. 387 (1947).
60. Roberts, *Jour. Res., NBS.*, **44**, p. 221 (1950).
61. Kuhn, *Kolloid-Z.*, **68**, p. 2 (1934); Kuhn and Kuhn, *Helv. chim. Acta*, **26**, p. 1394 (1943).
62. Huggins, *J. Phys. Chem.*, **43**, p. 439 (1939).
63. Debye, *Phys. Rev.*, **71**, p. 486 (1947).
64. Brinkman, *Appl. Sci. Res.*, **A1**, p. 27 (1947).
65. Huggins, *J. Appl. Phys.*, **10**, p. 700 (1939).
66. Alfrey, Bartovics and Mark, *J. Am. Chem. Soc.*, **64**, p. 1557 (1942).
67. Baker, *Ind. Eng. Chem.*, **41**, p. 511 (1949).
68. Boyer, *J. Appl. Phys.*, **20**, p. 540 (1949).
69. Zimm, *J. Chem. Phys.*, **16**, p. 1099 (1948); Outer, Carr and Zimm, *Ibid.*, **18**, p. 830 (1950).
69a. Kunst, *Rec. trav. chim.*, **69**, p. 125 (1950).
70. Cerf, *Compt. rend.*, **226**, p. 1586 (1948); *Ibid.*, **227**, p. 1221 (1948).
71. Kirkwood, *Rev. trav. chim. Pays-Bas*, **68**, p. 649 (1949).
72. Schmid and Rommel, *Z. Physik. Chem.*, **A85**, p. 97 (1939).
73. Schmid and Beuttenmüller, *Z. Elektrochem.*, **49**, p. 325 (1943); **50**, p. 209 (1944).
74. Natta and Baccaredda, *Gazz. chim. Ital.*, **79**, p. 364 (1949).

The Reliability of Telephone Traffic Load Measurements by Switch Counts

BY W. S. HAYWARD, JR.

(Manuscript received October 15, 1951)

The switch count method of telephone traffic measurement is subject to sampling errors. The nature of these errors is discussed and formulas are derived which describe the extent of the errors under normally encountered traffic conditions.

INTRODUCTION

Of prime importance to the telephone traffic engineer is the determination of the busy season busy hour load carried by groups of trunks or other circuits of a telephone switching system. Three direct methods of measuring such loads are found in the field today. These are:

a. Peg Count and Holding Time Method

The number of calls carried by the circuit group during the observation period is counted. This number multiplied by the average holding time per call (in hundreds of seconds) and divided by the length of the observation period (in hours) gives an estimate of the group load in units of hundred-call-seconds per hour (CCS). The major drawback to this peg count method is that it requires a separate determination of the average holding time per call for the group under observation. R. I. Wilkinson¹ has analyzed the sources of errors of holding time measurements. In addition, correlation between load and holding time introduces an error which has not been studied.

b. Switch Count Method

At fixed intervals the circuit group is scanned and the number of busy circuits is counted. The total number of busy conditions counted divided by the number of scans is, then, an estimate of the load on the group in units of average simultaneous calls or erlangs*. This estimate is generally converted to CCS (1 erlang = 36 CCS) by traffic engineers since the

* The name "erlang" for average simultaneous call was adopted at a plenary meeting of the CCIF at Montreux in October, 1946.

load entries of most traffic tables are in terms of CCS. For theoretical studies the erlang is a more convenient unit and will be retained here.

c. Continuous Method

The busy condition of each circuit is represented by a fixed increment of electrical current through an ampere-hour meter. The instantaneous current is then analogous to the calls simultaneously present so that the meter, which integrates the current, may be calibrated to indicate hundred-call-seconds or erlang-hours directly. Although this method is potentially the most accurate, practical difficulties have limited its use.

In addition to these direct methods, there are several methods of indirect load measurement which, relying more heavily on traffic theory, make use of partial load indications, such as duration of group busy or the number of calls finding the group busy. Such measurements are less reliable than the direct measurements particularly when applied to underloaded groups.

This paper is concerned with the reliability of switch count load measurements since this method appears to have prospects of considerably wider adoption in the future. Main emphasis will be placed, both qualitatively and by the application of error formulas, on the relative effects of various measurement and traffic parameters on the accuracy of switch count measurements. Where long derivations of formulas are required they are deferred to the Appendix.

SOURCES OF ERROR

As has been described, switch count measurements yield the average number of calls found present when a group of circuits is scanned at fixed intervals during an observation period. Usually only that period of the day during which the load is greatest is of interest to the traffic engineer. Because the load during such periods also fluctuates from day to day, measurements of the loads for several days must be averaged to provide a useful load estimate.

There are two main sources of error, therefore, in switch count estimates of telephone traffic loads:

1. Each individual count of busy circuits is separated from the next by a time interval during which changes in load are not detected. Consequently, the load indicated by measurement may differ appreciably from the actual load carried. This difference can be decreased by decreasing the interval between scans.

2. Even if the load carried during a measurement period were known very accurately, it is still only a sample of the many loads that might be offered by the same source of traffic under statistically identical conditions. Therefore, the average of several load readings may be expected to be somewhat in error as an estimate of the true average of the traffic source. The latter will be referred to as the *source* load to distinguish it from the *carried* load.

Mechanical and human errors are likely to be present as well but, since they are not inherent in the switch count method, they will be neglected here.

SWITCH COUNT ERROR

As shown in the Appendix, for periods of observation which are relatively long with respect to average holding time made on traffic with certain assumed characteristics, the average error of switch counts in estimating traffic load *carried* in the same period is zero. The coefficient of variation of the error, which is the standard deviation of the error expressed in per cent of the traffic load carried, is given by:

$$V_x \doteq 100 \sqrt{\left[r \operatorname{ctnh}\left(\frac{r}{2}\right) - 2 \right] \frac{\bar{t}}{a'NT}} \quad (1^*)$$

$$\operatorname{ctnh}\left(\frac{r}{2}\right) = \frac{1 + e^{-r}}{1 - e^{-r}} = \text{hyperbolic cotangent of } \frac{r}{2}$$

$$rc = T/\bar{t} > 20$$

where r = ratio of scan interval to holding time

\bar{t} = average holding time

a' = *carried* load in erlangs

c = number of switch counts

T = length of observation period

N = number of observation periods

and where the following assumptions are made:

- Calls originate individually and collectively at random.†
- Holding times are exponentially distributed.
- Congestion loss from the group is negligible.

* I have recently learned that these *carried* load formulas have been published by Conny Palm in *Tekniska Meddelanden från Kungl. Telegrafstyrelsen*, 1941. nr. 7-9.

† See T. C. Fry, *Probability and Its Engineering Uses*, D. van Nostrand Co. Inc., New York, p. 216, for a definition of this condition.

As shown in the Appendix, this formula simplifies, when $r \leq 2$, to

$$V_x \doteq \frac{100}{c \cdot T} \sqrt{\frac{1}{6a'N\bar{T}\bar{t}}} \quad (2)$$

where $c \cdot T$ = rate of scan in cycles per time unit. From equation (2) it is apparent that if the scan interval is of the order of a holding time, the error of an estimate of traffic *carried* is inversely proportional to the rate of scan and inversely proportional to the square root of average load, holding time and hours of observation. For example, take the case where switch counts are made during the busy hour, five minutes apart on a trunk group carrying calls with an average holding time of 3 minutes and an average load of 5 erlangs (180 CCS). What is the error in the estimated load *carried* if the readings for ten days are averaged? (Assume conditions (a), (b) and (c) are met.) We have

$$N = 10 \text{ observation periods}$$

$$T = 1 \text{ hour}$$

$$\bar{t} = 1/20 \text{ hour}$$

$$a' = 5 \text{ erlangs}$$

$$c = 12 \text{ scans per observation period}$$

$$rc = T/\bar{t} = 20 \text{ average holding times per observation period}$$

From equation (2) since $T/\bar{t} = 20$ and $r = T/\bar{c}\bar{t} = 1.7$

$$V_x = \frac{100}{12} \sqrt{\frac{1}{6 \cdot 5 \cdot 10 \cdot 1 \cdot 1/20}} = 2.15\%$$

If, as proposed in the Appendix, it is assumed that the error has a normal distribution, there is 90 per cent assurance that observed values will fall within $1.64V_x$, or in the example within 3.52 per cent, of the true average*. Note that this error limit would be halved if the rate of scan were doubled or if four times as many hours of observation were taken.

The coefficient of variation of the switch count error for constant values of T/\bar{t} as a function of r is plotted on Fig. 1 for one observation period of a one erlang load. For loads other than one erlang the coefficient of variation is found by dividing by $\sqrt{a'N}$. Thus in the example we have, using the dotted curve,

$$V_x = \frac{15}{\sqrt{5 \cdot 10}} = 2.1\%$$

* This assumes that a sufficient number of observations are taken so that a priori information may be neglected in making an estimate of the universe.

or more accurately using the solid curve,

$$V_x = \frac{16}{\sqrt{5 \cdot 10}} = 2.3\%$$

The error of using equation (2) is seen to be negligible for most purposes even when T/t is less than 20. The probability of an observation occurring within a given number of standard deviations is widely published for the normal curve. A few values are given below:

σ	P_z Probability of exceeding $\pm \sigma$ or $\pm zV$
0.6745	0.50
1.44	0.85
1.64	0.90
2.00	0.9545
3.00	0.9973

Fig. 2 is a plot for 40 observations of measured load vs carried load. Each observation was made for a half hour period on a panel line finder group with switch counts made at the start and middle of the period.

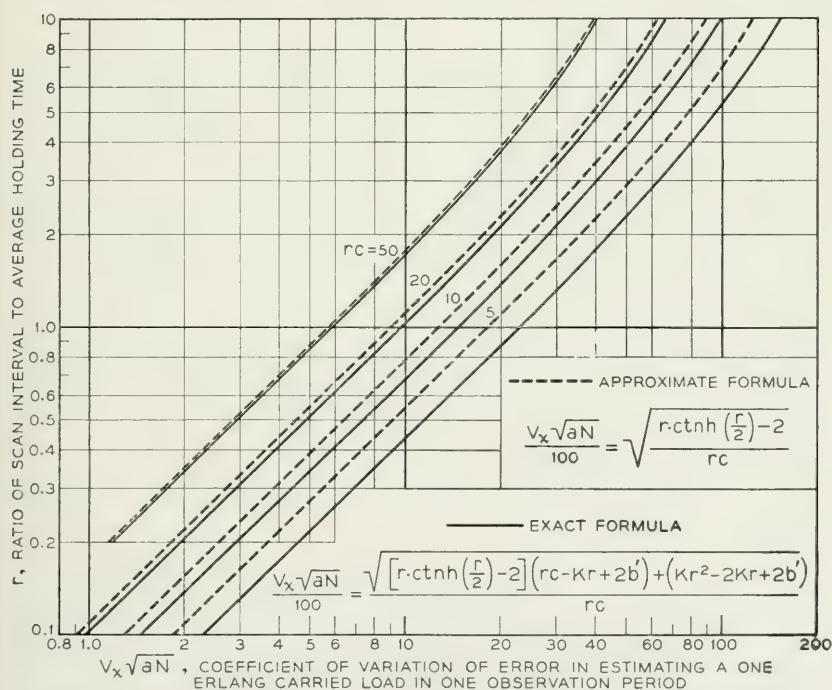


Fig. 1—Accuracy of switch count estimate of load actually carried.

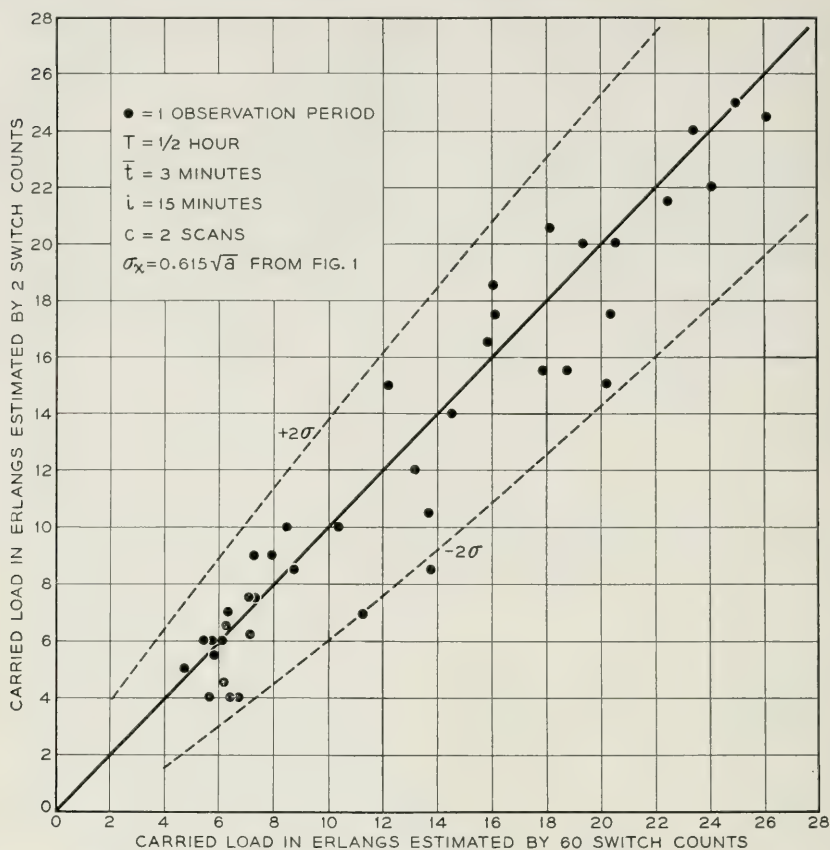


Fig. 2—Accuracy of switch count estimate of true average load.

This is compared with the average of switch counts made every 30 seconds which has a relatively negligible error. The average holding time per call for the group was 176 seconds. The accuracy of only two counts is surprisingly good and the observations are seen to lie satisfactorily between the 2σ limits.

ERROR OF TRAFFIC IN A GIVEN PERIOD AS AN ESTIMATE OF THE SOURCE LOAD

The average traffic carried in two different periods but generated by the same traffic source is subject to statistical variation. As a result, any measurement of load, even if measurement errors are eliminated, is only a sample of the wide range of traffic loads that might have been generated by the same source of traffic under identical circumstances.

J. Riordan has shown² that the standard deviation of the average traffic load for any one period is given by

$$\sigma_y = \sqrt{\frac{2a\bar{t}^2}{T^2} \left(\frac{T}{\bar{t}} - 1 + e^{-T/\bar{t}} \right)} \quad (3)$$

where a = the average source load

\bar{t} = average holding time per call

T = length of observation period

(Assumptions are as before with an additional one that all periods are in statistical equilibrium)

When $\bar{t}, T \ll 1$ this reduces to the form also given by F. W. Rabe³

$$\sigma_y = \sqrt{\frac{2a\bar{t}}{T}} \quad (4)$$

or expressed in a per cent of the average

$$V_y = 100 \sqrt{\frac{2\bar{t}}{aT}} \quad (5)$$

When N periods of length T are observed the coefficient of variation is reduced further to:

$$V_y = 100 \sqrt{\frac{2\bar{t}}{aNT}} \quad (6)$$

In the example of the previous section,

$$N = 10$$

$$T = 1$$

$$\bar{t} = 1/20$$

$$a = 5$$

$$V_y = \sqrt{\frac{2 \cdot 1/20}{5 \cdot 10 \cdot 1}} = 4.47\%$$

COMBINATION OF ERRORS

Evidently if switch count readings are used to estimate the average which may be expected in other periods, the two errors described above should both be taken into account. The errors are probably correlated but this correlation is weak and at present no method of allowing for it

is evident. Such a refinement would probably change the equation for standard deviation only slightly from that derived for the independent case; therefore independence will be assumed. The standard deviation of the sum of two independent variables is the square root of the sum of the squares of the component standard deviations:

$$\sigma_s = \sqrt{\sigma_x^2 + \sigma_y^2} \quad (7)$$

$$= \sqrt{\left[\operatorname{rectnh}\left(\frac{r}{2}\right) - 2 \right] \frac{\bar{t}a'}{NT} + \frac{2\bar{t}a}{NT}} \quad (8)$$

Assuming $\frac{a'}{a}$ is approximately unity, that is, that carried load is approximately equal to source load,

$$V_s = 100 \sqrt{\frac{\bar{t}}{anT} \operatorname{rectnh}\left(\frac{r}{2}\right)} \quad (9)$$

In the example given,

$$V_s = 4.96\%$$

There is, then, 90 per cent assurance that the source average is within $1.64 \times 4.96 = 8.1$ per cent of the observed average. Note that doubling the switch count rate (which halves the switch count error) reduces the total error only to 7.6 per cent (about 6.7 per cent improvement), while doubling the number of hours of observations reduces the error to 5.9 per cent (about 30 per cent improvement). Plots of the coefficient of variation of a one hour observation of a one erlang load versus scan rate for various average holding times are given in Fig. 3 for a wide range of holding times. The coefficient of variation of error in estimating other loads may be found from Fig. 3 by dividing the unit load coefficient by \sqrt{aNT} . In the example, the unit load coefficient is found, by entering Fig. 3 with $\bar{t} = 3$ minutes and rate of scan = c , $T = 12/1$ scan cycles per hour, to be 35.0 per cent. Dividing by $\sqrt{5 \cdot 1 \cdot 10}$ gives a coefficient of variation of 4.96 per cent as before. It is evident from Fig. 3 that increasing scan rates is not a universal way to improve the accuracy of *source* load estimates.

CHOICE OF SCAN RATES

What then governs the choice of scan rate? Evidently increasing the rate increases the accuracy of *carried* load estimates to any point de-

sired. This is far from true if *source* load is being estimated. If the cost of making a scan is constant, increasing the number of observation periods and decreasing the scan rate will improve accuracy of *source* load estimates without changing measurement costs. The number of hours available for measuring, of course, limits this procedure, while the increase in accuracy becomes negligible as r becomes large. On the other hand, if the cost of each observation is only slightly affected by the cost of making additional scans, a high scan rate might be justified.

In applying the above relationships to traffic measurements, the usual question raised by the traffic engineer will be either how many hours of data need he take to be reasonably sure of his estimate or, conversely, how sure is he of an estimate based on available data. Assuming as before that the error distribution is normal, the per cent plus or minus error limits within which a proportion, P_z , of the estimates will fall is given by zV_s ; the value of z corresponding to any selected P_z may be found from tables of the normal probability distribution. "Reasonably sure" is often taken to mean that there is 90 per cent assurance that the error does not exceed 5 per cent. When P_z is 0.90, z is 1.64, so that under this condition $1.64V_s = 0.05$, or $V_s = 0.0305$. Given scan rate and holding time, V_s is proportional to $1/\sqrt{aNT}$ according to equation (9) or Figure 3. When V_s is held constant, aNT is constant so that the plot of $\log NT$ against $\log a$ is linear, as shown in Figs. 4 and 5. The number of hours needed to meet any chosen reliability

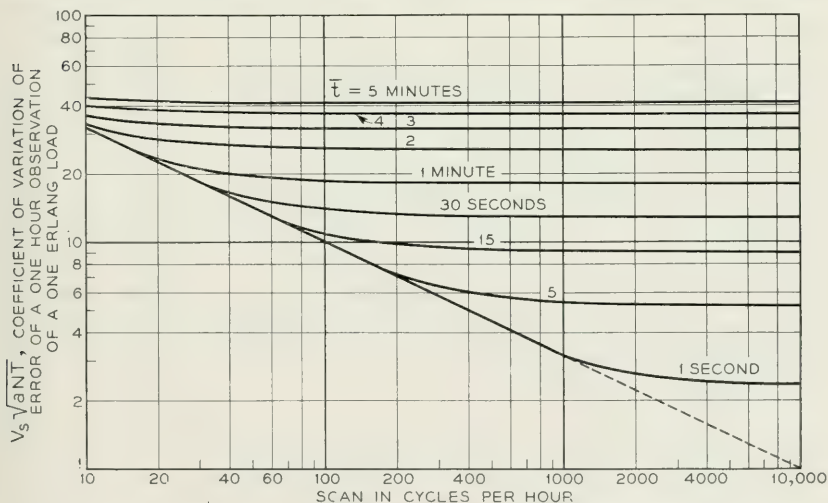


Fig. 3—Efficiency of switch counts for usage measurement.

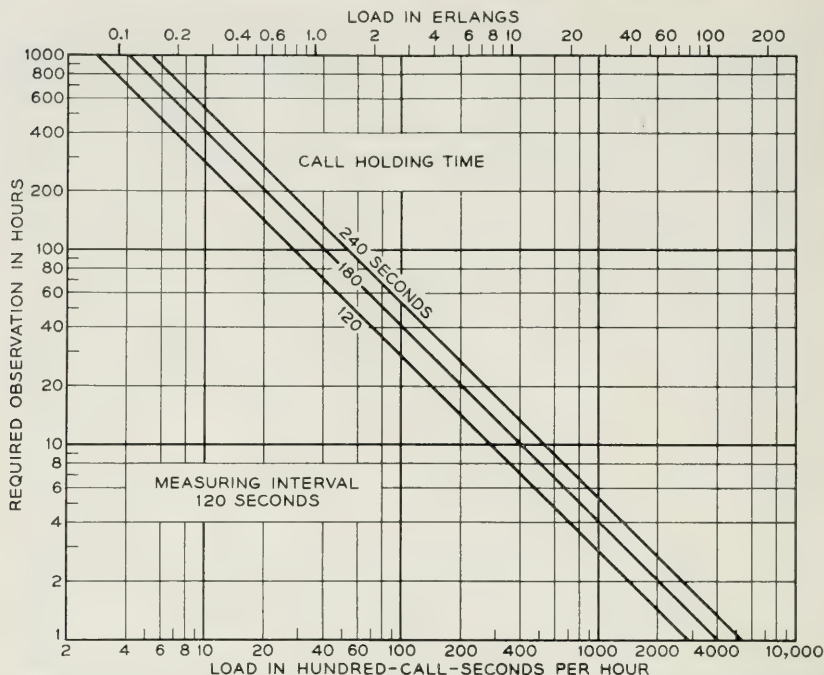


Fig. 4—Hours of measurement required for 90 per cent assurance that error in estimating *source* load does not exceed plus or minus 5 per cent when measuring interval is 120 seconds.

requirements may then be read directly from such graphs. In the second type of question, z , NT , scan rate and holding time are fixed so that zV_s is proportional to $1/\sqrt{a}$. Plotting $\log zV_s$ against $\log \sqrt{a}$ again gives a linear plot as shown on Fig. 6.

In the numerical example above, the limits of error corresponding to 90 per cent assurance may be read from Fig. 6 which is plotted for the appropriate assurance, average holding time and scan interval. Reading the error limits at the point where the 10 hours measured line crosses 180 CCS (5 erlangs) gives ± 8.1 per cent as before. Fig. 5 may be entered to find the total number of hours required to reduce this error to 5 per cent. Reading at the point where the 180 second holding time line crosses 180 CCS gives 26 hours.

QUALITATIVE EXTENSION OF THEORETICAL APPROACH

The original traffic assumptions made in deriving the theoretical results above are:

- a. Calls originate collectively and individually at random.

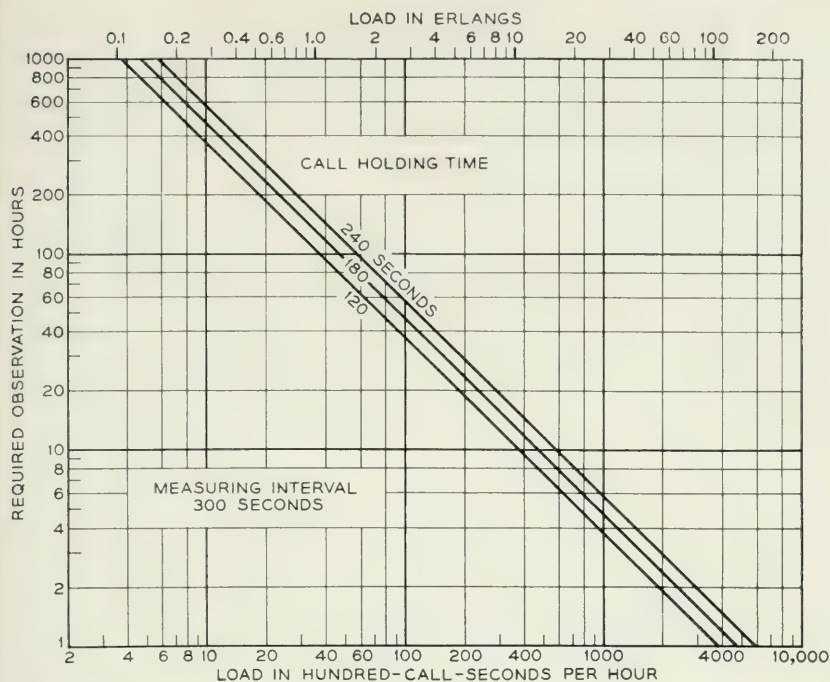


Fig. 5—Hours of measurement required for 90 per cent assurance that error in estimating *source* load does not exceed plus or minus 5 per cent when measuring interval is 300 seconds.

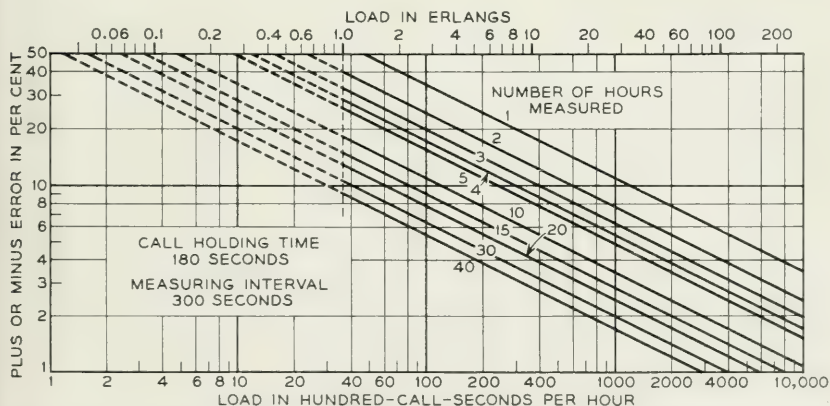


Fig. 6—Limits of error reached with 90 per cent assurance in estimating *source* load.

- b. Holding times are exponentially distributed.
- c. Congestion loss from the group is negligible.
- d. Observation periods are in statistical equilibrium.

How do departures from these assumptions affect the reliability of usage measurements?

a. Holding Time Distribution

Experience in application of delay and loss formulas has shown that theories based on exponential holding times are often applicable to other holding time distribution cases which have a wide range. However, for a constant holding time distribution special theories often are called for. The average and standard deviation of switch count estimates of *carried* load when holding time is constant, are given in part 2 of the Appendix. It is shown there that for estimates of *carried* load,

$$\bar{x} = 0$$

$$r \geq 1 \quad V_x = 100 \sqrt{\frac{\bar{t}}{a'NT}} (r - 1) \quad (10)$$

$$r \leq 1 \quad \text{minimum } V_x = 0 \quad (r = 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \text{ etc.})$$

$$\text{maximum } V_x = 100 \frac{r}{2} \sqrt{\frac{\bar{t}}{a'NT}} \quad (r = \frac{2}{3}, \frac{2}{5}, \frac{2}{7}, \text{ etc.}) \quad (11)$$

Since constant holding times found in practice are often very short, the case of $r \geq 1$ is the most likely to be met. For all values of r greater than one, the error given by formula (1) for exponential holding times is somewhat greater than the error given by formula (10) for constant holding times, so use of formula (1) for the constant holding time case is conservative. For values of r less than 1, the error is an oscillating function of r . The coefficient of variation varies from zero to 23 per cent above that for exponential holding times. Where r may not be accurately known the formula for exponential holding times again seems appropriate.

In making estimates of the *source* load when the holding time is constant, if $r \geq 1$, each scan is uncorrelated with any other, since no call can be counted twice, and may be considered a random sample of traffic. There are a total of Nc scans which have an average scan of a and standard deviation \sqrt{a} . The average error in estimating a is, therefore:

$$\bar{s} = 0$$

with coefficient of variation

$$V_s = 100 \sqrt{\frac{1}{aNc}} = 100 \sqrt{\frac{\bar{t}}{aNT}} r \quad (12)$$

Equation (12) may also be derived with the procedure used for equation (9) using $\sigma_s^2 = \sigma_x^2 + \sigma_y^2$. For values of r large enough to make $\text{ctnh}\left(\frac{r}{2}\right) \doteq 1$ equation (12) is approached by equation (9). For smaller values of r (but with r still greater than 1), V_s for constant holding times is less than V_s for exponential holding times. When $r = 1$, there is no *carried* load error. For values of r less than 1, the coefficient of variation of error in estimating *source* load average will vary from

$$\sqrt{\frac{\bar{t}}{aNT}} \quad \text{to} \quad \sqrt{\frac{\bar{t}}{aNT} \left(1 + \frac{r^2}{2}\right)}$$

depending on the exact value of r . It is interesting to note that V_s for $r = 0.5$ is the same as for $r = 1.125$.

b. Loss

The effect of loss in the group depends upon the disposition of the lost calls. In general, accuracy in measuring *carried* load increases with increased loss because under these circumstances fewer load changes occur between scans. This is evident in the extreme case of a group which is 100 per cent loaded; a single switch count gives a correct reading for any length period. Obviously load readings at 100 per cent occupancy are not very useful in estimating *offered* loads since the amount of lost load cannot even be guessed at. However, in the cases of lost calls held (Poisson) or cleared (Erlang B), the *offered* load may be estimated from the *carried* load (less and less accurately as occupancy increases) and in the case of lost calls delayed the *offered* and *carried* loads are likely to be the same even at high occupancies. With high loss, therefore, estimates of *source* load are subject to errors not considered in deriving equation (99); however, switch count error in estimating *carried* load will be materially less than predicted by equation (1).

c. Random Call Origination

On trunk groups which are alternate routes, calls may no longer be considered as originating at random. The resultant grouping of call originations will tend to decrease the accuracy of switch count measurements in estimating *carried* load; however, there is a corresponding decrease in accuracy in estimating the *source* load from the *carried* load so that accuracy in estimating *carried* load may be less worthwhile.

d. Statistical Equilibrium

Statistical equilibrium may be thought of as the absence of trends in subscriber calling rates or holding times with the passage of time. The effect of trends on switch count accuracy in measuring *carried load* is very small except where the changes in traffic level are frequent and abrupt with respect to the scan frequency. Such traffic behavior is rare.

Trends within the busy hour complicate the problem of estimating the average *source load*. However, it can be shown that if the trends are small (in the order of 10 per cent to 20 per cent) little error is introduced by assuming that no trend exists. Large trends (in the order of 100 per cent), however, may indicate that the traffic source is so unstable that more hours of traffic data should be taken in order to insure that the sample is representative.

Trends from day to day do not affect the *source* load estimates in the same way as within hour trends. The *source* loads are seldom exactly the same on any two days although in most offices a load pattern is repeated from week to week. The traffic engineer may be interested in the average *source* load of either a typical week day in the busy season or, sometimes, of the average of the two highest days in the week. As long as the *source* load of each particular day remains close to the average for that day of the week, the general average for several different days of the week, will be known with about the same accuracy as if they had all come from a common source. If, however, there is no stable pattern in the *source* load, a third error in estimating the average is generated. There is some difficulty in determining whether or not variations in load, as indicated by measurements, are due to sampling variations or to an unstable source. Quality control methods might be used to detect instability but gathering and processing sufficient data for such an analysis might prove uneconomical. In general, if a traffic engineer feels that his *source* load is unstable he will need more hours of data than indicated by formula (9) to meet a given criterion of reliability.

CONCLUSIONS

A theoretical approach to the problem of the accuracy of switch count measurements in estimating *carried load* and average *source load* has been explored. It is believed that the assumptions made are satisfied sufficiently often in practice to enable fairly wide application of the results of this exploration to traffic measurements. However, it should be kept in mind that where the assumptions are clearly not valid, special allowances will need to be made. In any case, the confidence placed in

usage measurements by a traffic engineer is a function of his experience and judgment. It is hoped that the results of this study will add to the knowledge essential to sound traffic engineering.

APPENDIX

DERIVATION OF SWITCH COUNT ERROR IN ESTIMATING CARRIED LOAD— WITH EXPONENTIAL HOLDING TIMES

This derivation is based on a similar derivation by R. I. Wilkinson¹. However, since load rather than holding time is of interest here, the emphasis has been somewhat shifted.

Assume that switch count measurements are being taken on traffic with:

- a. Calls originated individually and collectively at random
- b. Exponentially distributed holding times
- c. Negligible loss

Let i = interval between scans

\bar{t} = average holding time

a' = traffic carried, in erlangs

T = length of observation period

$r = \frac{i}{\bar{t}}$ = number of holding times in a scan interval

$c = \frac{T}{\bar{t}}$ = number scans in observation period

$rc = \frac{T}{i}$ = number of holding times in observation period

N = number of observation periods.

Consider that the observation period begins with the first scan and ends i time units after the last scan. It is desired to find the error in estimating the true load carried by averaging the number of circuits found busy on each scan. Following Wilkinson's method we will first estimate the error of the switch count method in measuring the contribution of a single call to total usage and then modify it to take account of n calls. Calls of two types must be considered, those originating outside the interval and extending into it, Type I, and those originating within the interval, Type II. Both types may be subdivided depending on whether or not they extend beyond the end of the observation period. These are

indicated in Fig. 7. Only that part of a call which falls within the observation period contributes to the usage of that period. First the error made by switch counts in measuring this contribution will be derived.

Type I

Consider a call which is already in progress at the start of the observation period. Its duration beyond that point, according to theory, will be exponentially distributed about an average of \bar{t} .

If this duration, t , is between 0 and i , the call will be counted once (a measured contribution of i erlang hours) and a positive error of $x = i - t$ will be made. The same error will be made if $t = 2i - x$ so that the call

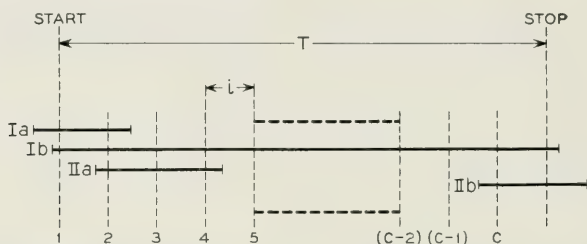


Fig. 7—Graphical indication of the two types of calls with their two subdivisions.

is counted twice and so forth. Summing all the ways of making an error x , we have:

$$P(x) dx = f(i - x) + f(2i - x) + \cdots f(ci - x) \quad (1)$$

where $f(i - x)$ is the probability of $t = i - x$ and

$$f(t) = \frac{1}{\bar{t}} e^{-t/\bar{t}} dx$$

Calls lasting beyond ci neither start nor end in the observation period so that their contribution is measured without error. For these:

$$P(0) = P(t \geq ci) = e^{-rc} \quad (2)$$

Therefore:

$$P_{x>0}(x) dx = \frac{1}{\bar{t}} e^{-\frac{i-x}{\bar{t}}} dx + \frac{1}{\bar{t}} e^{-\frac{2i-x}{\bar{t}}} dx + \cdots + \frac{1}{\bar{t}} e^{-\frac{ci-x}{\bar{t}}} dx \quad (3)$$

Letting

$$y = \frac{x}{t}; \quad e^{-cr} = b = 1 - b'; \quad K = \sum_{n=0}^{c-1} e^{-nr} = \frac{b'}{1 - e^{-r}} \\ P_{x>0}(x) dx = e^y e^{-r} K dy \quad (4)$$

The moment generating function $M_I(\alpha)$ of y is:

$$M_I(x) = \int_0^r P(y) dy e^{\alpha y} + e^{-rc} \\ = b + K \frac{e^{r\alpha} - e^{-r}}{1 + \alpha} \quad (5)$$

Neglecting terms of order higher than α^2 ,

$$M_I(\alpha) = 1 + \alpha(rK - b') + \frac{\alpha^2}{2} (Kr^2 + 2b' - 2rK) \quad (6)$$

Type II

Calls of Type II may have either positive or negative errors given by:

$$P_{x \leq 0}(x) dx = \frac{i+x}{i} [f_0(-x) + f_1(i-x) + f_2(2i-x) \\ + \cdots + f_{c-1}((c-1)i-x)] \quad (7)$$

$$+ g_0(-x) + g_1(i-x) + g_2(2i-x) + \cdots + g_{c-1}[(c-1)i-x]$$

$$P_{x \geq 0}(x) dx = \frac{i-x}{i} [f_1(i-x) + f_2(2i-x) + \cdots + f_c(ci-x)]$$

where $f_n(ni-x)$ = probability that a Type II call has length $ni-x$ and ends before the end of the observation period.

$$= \frac{1}{t} e^{-\frac{n(i-x)}{t}} \cdot \frac{c-n}{c} dx$$

$g_n(ni-x)$ = probability that a Type II call starts $ni-x$ before the end of the observation period and ends after the end of the observation period.

$$= \frac{1}{T} e^{-\frac{n(i-x)}{t}} dx$$

Equation (7) becomes:

$$P_{x \leq 0}(x) dx = \left[\frac{i+x}{i} \sum_{n=0}^{c-1} \frac{1}{\bar{t}} e^{-\frac{n i-x}{\bar{t}}} \cdot \frac{c-n}{c} + \sum_{n=0}^{c-1} \frac{1}{\bar{T}} e^{-\frac{n i-x}{\bar{t}}} \right] dx$$

$$P_{x \geq 0}(x) dx = \left[\frac{i-x}{i} \sum_{n=1}^c \frac{1}{\bar{t}} e^{-\frac{n i-x}{\bar{t}}} \cdot \frac{c-n}{c} \right] dx \quad (8)$$

Letting

$$\frac{x}{\bar{t}} = y, \quad K = \sum_{n=0}^c e^{-nr} \text{ as before}$$

and noting that

$$\sum_{n=0}^{c-1} n e^{-nr} = \frac{K - cb}{1 - e^{-r}}$$

$$P_{x \leq 0}(x) dx = e^y \left[\left(1 + \frac{y}{r} \right) \left(K - \frac{1}{c} \frac{e^{-r} K - cb}{1 - e^{-r}} \right) + \frac{K}{rc} \right] dy$$

$$P_{x \geq 0}(x) dx = e^{-r} e^y \left[\left(1 - \frac{y}{r} \right) \left(K - \frac{1}{c} \frac{K - cb}{1 - e^{-r}} \right) \right] dy \quad (9)$$

The moment generating function of this pair of equations is the sum of their separate m.g.f.'s:

$$M_H(\alpha) = \int_{-r}^0 P_{y \leq 0}(y) e^{\alpha y} dy + \int_0^r P_{y \geq 0}(y) dy e^{\alpha y} dy$$

$$rc + K + c - 2 \frac{c - K e^{-r}}{1 - e^{-r}} + \frac{c - K}{1 - e^{-r}} e^{\alpha r} + \frac{(c - K) e^{-r}}{1 - e^{-r}} e^{-\alpha r} \quad (10)$$

$$= \frac{\quad}{rc(1 + \alpha)^2}$$

Neglecting terms of order higher than α^2 ,

$$M_H(\alpha)$$

$$= \frac{1}{rc} \left\{ rc + \alpha(b' - rK) + \frac{\alpha^2}{2} \left[\left(r \operatorname{ctnh} \left(\frac{\alpha}{2} \right) - 2 \right) (rc - rK + 2b') \right] \right\} \quad (11)$$

Now the number of Type I calls present in an observation is a variable—with average “ a ” and a Poisson distribution. Similarly the number of Type II calls is a variable, independent of the number of Type I calls, with an average of “ $a \frac{T}{\bar{t}}$ ” or “ arc ” and a Poisson distribution. Ac-

According to the laws governing the compounding of variables the moment generating function of the sum of n variables y , when n is also variable with generating function $G(t)$, is $G(M(\alpha))$ where $M(\alpha)$ is the moment generating function of y .

The generating function of a Poisson variable with average " a " is $e^{-a+a t}$ so that

$$\begin{aligned} G(M_I(\alpha)) &= e^{-a+a M_I(\alpha)} \\ G(M_{II}(\alpha)) &= e^{-arc+arc M_{II}(\alpha)} \end{aligned} \quad (12)$$

These independent variables may be added by multiplying their moment generating functions to give the m.g.f. of the total measurement error of the *carried* load

$$M(\alpha) = e^{-(a+arc)+a M_I(\alpha)+arc M_{II}(\alpha)} \quad (13)$$

From (6), (11) and (13) the following parameters are found:

$$\bar{y} = 0 \quad (14)$$

$$\sigma_y^2 = arc \left[\left(r \operatorname{ctnh} \left(\frac{r}{2} \right) - 2 \right) \left(1 - \frac{K}{c} + \frac{2b'}{rc} \right) + \left(\frac{rK}{c} - 2 \frac{K}{c} + \frac{2b'}{rc} \right) \right]$$

If, now, rc is sufficiently large

$$\sigma_y \doteq \sqrt{arc \left[r \operatorname{ctnh} \left(\frac{r}{2} \right) - 2 \right]} \quad (15)$$

It is more convenient to deal with the standard deviation expressed as per cent of the *carried* erlang load, the coefficient of variation. This is done by multiplying both sides of equation (15) by \bar{t} to convert the time dimension from holding times to hours, dividing by T to convert from erlang-hours to erlangs, dividing by a' to convert to proportion of *carried* load, and multiplying by 100 to convert to per cent. Assuming $\frac{a}{a'^2}$ is approximately $\frac{1}{a'}$:

$$V_x \doteq 100 \sqrt{\frac{\left[r \operatorname{ctnh} \left(\frac{r}{2} \right) - 2 \right] \bar{t}}{a' T}}$$

When N observations are made this reduces further to

$$V_x \doteq 100 \sqrt{\frac{\bar{t}}{a' N T} \left[r \operatorname{ctnh} \left(\frac{r}{2} \right) - 2 \right]}$$

$$\text{Now } \operatorname{ctnh}(x) = \frac{1}{x} + \frac{x}{3} - \frac{x^3}{45} + \frac{2x^5}{945} - \dots \quad (x^2 < \pi^2)$$

and for $r \leq 2$

$$\frac{r/2}{3} \gg \frac{(1/2)^3}{45}$$

Therefore

$$\operatorname{retnh}\left(\frac{r}{2}\right) \doteq 2 + \frac{r^2}{6}$$

$$V_x \doteq 100 \sqrt{\frac{\bar{t}}{a'NT}} \frac{r^2}{6} = \frac{100}{c/T} \sqrt{\frac{1}{6a'NT\bar{t}}} \quad (r \leq 2) \quad (16)$$

The error in *carried* load may be considered as the sum of a large number of independent errors. Its distribution may, therefore, be expected to approach the normal distribution. Comparison of the third and fourth moments of the normal distribution with those of the error distribution (which may be obtained from equation (13)) show good agreement for values of a' greater than 1.

DERIVATION OF SWITCH COUNT ERROR IN ESTIMATING CARRIED LOAD WITH CONSTANT HOLDING TIMES

Wilkinson has shown¹ that, for constant holding time, switch count error in measuring the holding time of one call has an average

$$\bar{x} = 0$$

and standard deviation

$$\sigma_x = \sqrt{-x_1 x_2}$$

where $T \gg \bar{t}$

$$T \gg i$$

x_1 = negative error

x_2 = positive error

Divide the problem into two parts:

1. For $r > 1$

$$x_1 = -\bar{t}$$

$$x_2 = i - \bar{t}$$

$$\sigma_x = \sqrt{\bar{t}i - \bar{t}^2} = \bar{t}\sqrt{r - 1} \quad (17)$$

2. For $r \leq 1$

$$\text{Min. } \sigma_x = 0 \quad \text{for} \quad r = 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \text{ etc.}$$

$$\text{Max. } \sigma_x = \sqrt{\frac{\bar{t}}{2} - \frac{\bar{t}}{2}} = \frac{\bar{t}}{2} = \bar{t} \frac{r}{2} \quad (18)$$

$$\text{for} \quad r = \frac{2}{3}, \frac{2}{5}, \frac{2}{7}, \text{ etc.}$$

Expressing this error in terms of carried load and proceeding as in Part I of the Appendix

$$1. \ r > 1 \quad V_x = 100 \sqrt{\frac{\bar{t}}{a'NT}} (r - 1) \quad (19)$$

$$2. \ r \leq 1 \text{ Min. } V_x = 0$$

$$\begin{aligned} \text{Max. } V_x &= 100 \frac{r}{2} \sqrt{\frac{\bar{t}}{a'NT}} \\ &= \frac{100}{c/T} \sqrt{\frac{1}{4a'NT\bar{t}}} \end{aligned} \quad (20)$$

Equation (20) compares favorably with the exponential holding time coefficient of variation of error of

$$\frac{100}{c/T} \sqrt{\frac{1}{6a'NT\bar{t}}}$$

REFERENCES

1. R. I. Wilkinson, "The Reliability of Holding Time Measurements," *Bell System Tech. J.*, **20**, pp. 365-404, October, 1941.
 2. J. Riordan, "Telephone Traffic Time Averages," *Bell System Tech. J.*, **30**, pp. 1129-1144, October, 1951.
 3. F. W. Rabe, "Variations of Telephone Traffic," *Electrical Communications*, **26**, pp. 243-248, 1948.
- The following books contain descriptions of the use of generating functions in solving probability problems:
4. A. C. Aitken, *Statistical Mathematics*, Oliver and Boyd, Ltd., Edinburgh, 1947, pp. 16-23.
 5. W. Feller, *An Introduction to Probability Theory and Its Applications*, John Wiley & Sons, Inc., New York, 1950, Chap. 11.

Network Representation of Transcendental Impedance Functions

BY M. K. ZINN

(Manuscript received November 5, 1951)

The purpose of the paper is to show that the admittance or impedance of certain continuous structures, such as, for example, a finite length of transmission line of any sort, or resonant cavity, can be represented exactly at all frequencies by a network comprising lumps of constant resistance R , inductance L , conductance G and capacitance C . The network will contain an infinite number of branches, in general, although a finite number may be used if it is desired to represent only certain modes.

The procedure is based upon a proposition known to students of function theory as "Mittag-Leffler's theorem," which amounts, roughly, to an extension of rational functions to apply to transcendental functions of the type encountered in the theory of continuous structures.

Several illustrative examples of the network synthesis are given.

GENERAL

Students of network theory are familiar with the fact that the impedance at a pair of terminals in a linear network comprising a finite number of resistors, inductors and capacitors, connected in any manner, is a rational function of the frequency having, in general, the fractional form of one polynomial divided by another. They are also familiar with the partial fraction rule whereby the function can be broken up into a series of elementary fractions, each of which exhibits one of the poles of the original function. This form is sometimes useful in the problem of network synthesis, where the impedance function is given and the object is to find a network having this impedance.

The purpose of the present paper is to show how a similar procedure can be carried out for certain transcendental impedance functions pertaining to structures having distributed constants, such as, for example, a resonant cavity or a piece of transmission line. The method employs a well-known proposition of function theory, which is usually referred to as Mittag-Leffler's theorem. This theorem provides a tool for breaking up a transcendental meromorphic function into an infinite series of simple fractions in much the same way as the partial fraction rule is used to break up a rational meromorphic function. The series representation

provides a means of determining a network of resistors, inductors and capacitors that will have an impedance equal to the specified transcendental impedance function. This process will be referred to as obtaining a "network representation" of the function. If the given function is the impedance of some continuous (i.e., non-lumped) electric structure, the result will be an equivalent network for the structure. For other purposes, such as, possibly, analogue methods of computing, the given function may not arise from any electrical structure. In either case, the network representations to be derived are possible only if the function satisfies certain restrictions, which are stated in the section immediately following.

The discussion is confined to transcendental impedance functions because of the technological interest in the electromagnetic structures with which they are associated and because they have not received as much attention as rational functions in the literature dealing with network synthesis. The problem with which this paper is concerned can then be stated as follows: given, a transcendental impedance function satisfying certain conditions: to determine a network comprising elements of constant resistance, inductance and capacitance whose driving-point impedance function, at a pair of terminals, will equal the given function at all frequencies, real and complex (except at the poles).

For illustration of the procedure, three examples are given. The first is the impedance of a short-circuited or open-circuited transmission line in which the distributed primary constants, R , L , G and C are assumed to be invariable with frequency. The second and third examples are the impedances of resonant cavities driven in two different modes. In these examples the variation of resistance with frequency, due to "skin-effect," is taken into account.

IMPEDANCE FUNCTIONS

The functions under discussion will be referred to as "impedance functions" with the understanding that the term is meant to include "admittance functions" as well. By reason of the duality principle that runs through all electric circuit theory, any general proposition developed for one must apply to the other. The functional designation, $F(p)$, will be used to denote either an impedance or an admittance function. When a distinction is necessary, the impedance will be designated by $Z(p)$ and the admittance by $Y(p)$. The independent complex variable p is the generalized radian frequency. (For sustained sinusoidal currents and voltages, $p = i\omega = 2\pi if$ where f is the real frequency.)

For the applications contemplated, $F(p)$ is a transcendental mero-

morphic function, which term implies that the function is given by the ratio of two entire functions, one or both of which is transcendental, and that the singularities of the function are ordinary poles, except for the point at infinity, which is an essentially singular point. In order to realize the particular network developments to be given, it will be supposed that the function satisfies the further restrictions given below:

(1) All the poles lie in the left half of the p -plane with none on the imaginary axis.

(2) $F(\bar{p}) = \bar{F}(p)$. (The superbar denotes the complex conjugate of the unbarred symbol.)

(3) Real part $[F(i\omega)] \geq 0$ for all real values of ω .

These three conditions are necessary to insure that the function is the impedance of a possible linear, passive electric circuit structure. Interpreted physically in terms of this possible equivalent structure, the first condition specifies that the structure shall be stable; that is, every natural mode of oscillation dies away exponentially. The second condition specifies that the natural oscillations are real functions of time. The third condition specifies that if a sinusoidal current flows at the driving-point terminals of the equivalent structure, the average real power delivered to it will be positive. Since these three conditions, or their equivalents, are frequently mentioned in discussions of network theory, it is assumed that they are understood without more detailed explanation.

In addition to the above restrictions on the form of the impedance function, the following two conditions, while not necessary, will be imposed to limit the scope of the discussion:

(4) All the poles of $F(p)$ are simple.

(5) $F(p) = 0(1)$, exactly, as $|p| \rightarrow \infty$ everywhere except at the poles.

Condition (4), while limiting the scope of the exposition required, does not restrict the application of the results in any important way, because most impedance functions for which a network representation may be required have only simple poles.

Condition (5) implies that as p increases along any straight line drawn through the origin and not passing through any pole of $F(p)$, the modulus of $F(p)$ either approaches a limit or oscillates between finite limits. The physical implication of this condition is that the response of the network as a function of time to a suddenly applied cause begins with a discontinuity of the same degree as that of the cause. For example, the current response of the network to an applied step of voltage begins with a finite discontinuity. This behavior is a characteristic of continuous (non-lumped) electromagnetic structures, which furnish the principal application of the network developments to be described.

MITTAG-LEFFLER'S THEOREM⁷

Let the poles of the given function $F(p)$ be $p_1, p_2, p_3 \dots$, where

$$0 < |p_1| \leq |p_2| \leq |p_3| \dots$$

and let the residues at the poles be $A_1, A_2, A_3 \dots$, respectively. Suppose that it is possible to draw a sequence of closed contours, C_n , such that C_n encloses $p_1, p_2, \dots p_n$ but no other poles and such that the minimum distance of C_n from the origin tends to infinity with n . Suppose also that $F(p)$ satisfies conditions (2), (4) and (5) above. Then Mittag-Leffler's theorem gives the following series development for $F(p)$:

$$F(p) = F(0) + \lim_{N \rightarrow \infty} \sum_{n=-N}^N \left(\frac{A_n}{p - p_n} + \frac{A_n}{p_n} \right) \quad (1)$$

The notation here used employs the convention that

$$p_{-n} = \bar{p}_n \quad \text{and} \quad A_{-n} = \bar{A}_n,$$

since, by virtue of condition (2), the poles occur in conjugate complex pairs. The value, $n = 0$, then allows for a pole on the negative real axis.

Given any suitable function, the procedure is to determine its value for $p = 0$ and the location of its poles. The residues are next determined by

$$A_n = \lim_{p \rightarrow p_n} (p - p_n)F(p).$$

Then the Mittag-Leffler expansion can be written down at once.

NETWORK REPRESENTATION

In the series (1) the terms occur in pairs with conjugate complex poles and residues. The object is to obtain a network representation of each such pair of terms. If $F(p)$ is taken as an admittance, the branches representing the pairs of terms will all be connected in parallel; if $F(p)$ is taken as an impedance, they will all be connected in series.

Methods for obtaining a network representation for a rational function, such as the one comprising a pair of terms in the series (1), are well known. It is only necessary to describe certain procedures of particular application to the present problem. Brune⁵ has stated that the necessary and sufficient condition for a network representation of a rational function of p to be realizable is that it be a "positive real function," that is, a function that is real for real values of p and whose real part is positive,

or zero, when the real part of p is positive, or zero. In view of conditions (1) and (2) above, only one test¹² need be applied to each pair of terms of the series (1): the sum of a pair of terms will be a positive real function if, and only if, the real part of their sum is greater than, or equal to, zero for all purely imaginary values of p .

The general term pair for which a network representation is sought is

$$F_n(p) = \frac{A_n}{p - p_n} + \frac{\bar{A}_n}{p - \bar{p}_n} + \frac{A_n}{p_n} + \frac{\bar{A}_n}{\bar{p}_n} = P_n(p) - P_n(0) \quad (2)$$

Evidently two cases can be distinguished at the outset, depending upon whether $P_n(0)$ is positive or negative. If $P_n(0)$ is positive, the network branch, in order to be realizable, should be designed to represent $P_n(p)$. The left-over negative term, $-P_n(0)$, then can be absorbed in the positive first term, $F(0)$, of the series (1); more will be said of this later. If, on the other hand, $P_n(0)$ is negative, the network branch should represent the whole term, $P_n(p) - P_n(0)$. This procedure insures that the real part of the branch impedance will be positive, or zero, at zero and infinite frequencies. To guarantee that the resistance is positive at all other frequencies requires further tests now to be specified.

Let the real and imaginary coefficients of the poles and residues of the n^{th} term be

$$\begin{aligned} p_n &= -\alpha_n + i\beta_n, & \bar{p}_n &= -\alpha_n - i\beta_n \\ A_n &= a_n + ib_n, & \bar{A}_n &= a_n - ib_n \end{aligned}$$

(With this notation, α_n and β_n are always positive; a_n and b_n can be either positive or negative.) Then (dropping the subscripts)

$$\begin{aligned} P(p) &= \frac{2(a\alpha - b\beta) + 2ap}{\alpha^2 + \beta^2 + 2\alpha p + p^2} \\ R[P(i\omega)] &= \frac{2(a\alpha - b\beta)(\alpha^2 + \beta^2) + 2\omega^2(a\alpha + b\beta)}{(\alpha^2 + \beta^2)^2 + 2\omega^2(\alpha^2 - \beta^2) + \omega^4} \\ P(0) &= \frac{2(a\alpha - b\beta)}{\alpha^2 + \beta^2} \\ R[P(i\omega) - P(0)] &= \frac{-2(a\alpha^3 - 3\alpha^2b\beta - 3a\alpha\beta^2 + b\beta^3)\omega^2 - 2(a\alpha - b\beta)\omega^4}{(\alpha^2 + \beta^2)[(\alpha^2 + \beta^2)^2 + 2(\alpha^2 - \beta^2)\omega^2 + \omega^4]} \end{aligned} \quad (3)$$

The necessary and sufficient conditions¹² for the real part of a rational function of p to be positive, or zero, for purely imaginary values of p are that the function be positive for $p \rightarrow \pm i\infty$ and have no imaginary roots of odd multiplicity. When this test is applied to the functions $P(p)$ and

$P(p) - P(0)$, as given by (3), the following conditions are obtained: $P(p)$ will be a positive real function if, and only if,

$$a\alpha - b\beta > 0; \quad \text{i.e.} \quad P(0) > 0 \quad (4)$$

and

$$a\alpha + b\beta > 0$$

$P(p) - P(0)$ will be a positive real function if, and only if,

$$a\alpha - b\beta < 0; \quad \text{i.e.} \quad P(0) < 0 \quad (5)$$

and

$$a\alpha^3 - 3\alpha^2b\beta - 3a\alpha\beta^2 + b\beta^3 < 0.$$

If all terms of the series satisfy one or the other of these conditions, network branches can be devised to represent all the terms and all the R, L, G, C elements of the branches will be positive.

In case all the terms are of the type where $P_n(0)$ is positive, so that the network branches are made to represent $P_n(p)$, the left-over constant terms can be collected and added to the first term, $F(0)$, of the series. This collection of terms then must be represented by a final branch of pure resistance, or conductance, of value,

$$F(0) - \sum_{n=0}^{\infty} P_n(0)$$

If the sum of the variable terms approaches zero for $p \rightarrow \pm i\infty$, the final constant term supplies the high frequency resistance of the function $F(p)$ and since this must be positive, if condition (3) is satisfied, the final resistive element will be positive. If the series converges non-uniformly, the sum of the variable terms can have a value other than zero as $p \rightarrow \pm i\infty$ in spite of the fact that every term approaches zero individually. In that case (see example 1) all or part of the high frequency resistance may be supplied by the sum of the variable terms.

In case all the terms are of the type where $P_n(0)$ is negative, so that the network branches are made to represent the sum, $P_n(p) - P_n(0)$, of the variable and constant terms and the series is uniformly convergent, all the high frequency resistance is provided by the branches representing these terms. The first term, $F(0)$ then supplies the dc resistance, which is positive by condition (3). Non-uniform convergence can modify this division of high- and low-frequency resistance, however.

Cases can arise in which the series contains terms of both types. In such a case the dc resistance, or high frequency resistance, or both, of

the given function might be less than the sum of the variable terms for these frequencies, with the result that the final resistance branch would be negative for either the series or parallel type of network development.

To make the procedure as concrete as possible, particular forms of networks are described in the section following with explicit formulas for computing their elements.

NETWORK FORMULAS

Simple forms of network branches are shown in Figs. 1 and 2. Those of Fig. 1, referred to as branches of "the first kind" are suitable for connection in parallel where the given function $F(p)$ is an admittance, $Y(p)$, while networks of "the second kind," shown in Fig. 2, are suitable for connection in series to represent an impedance, $F(p) = Z(p)$. The networks of Figs. 1a and 2a apply where the value $P_n(0)$ of the general term is positive, while Figs. 1b and 2b apply where $P_n(0)$ is negative. Figs. 3 and 4 illustrate, respectively, networks of the types of Figs. 1a and 2a

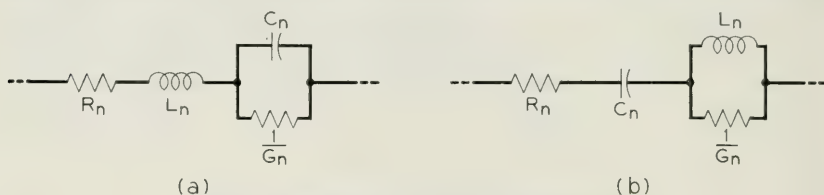


Fig. 1—General branches of the first kind.

<i>Fig. 1a</i>	<i>Fig. 1b</i>	
(use where $F(p) = Y(p)$ and $Y_n(0) > 0$)	(use where $F(p) = Y(p)$ and $Y_n(0) < 0$)	
$L_n = \frac{1}{2a_n}$	$L_n = \frac{\beta_n^2(\alpha_n^2 + \beta_n^2)^2(a_n^2 + b_n^2)}{2M^3}$	
$\frac{1}{L_n C_n} = \beta_n^2 \left(\frac{b_n^2}{a_n^2} + 1 \right)$	$\frac{1}{L_n C_n} = \frac{M^2}{\beta_n^2(a_n^2 + b_n^2)}$	
$\frac{G_n}{C_n} = \frac{1}{a_n} (a_n \alpha_n - b_n \beta_n)$	$G_n L_n = -\frac{a_n \alpha_n - b_n \beta_n}{M}$	(6)
$\frac{R_n}{L_n} = \frac{1}{a_n} (a_n \alpha_n + b_n \beta_n)$	$R_n C_n = \frac{N}{M(\alpha_n^2 + \beta_n^2)}$	
$G_o = Y(0) - \sum_{n=0}^{\infty} Y_n(0)$	$G_o = Y(0)$	

connected to form the completed network with the final non-reactive branch, G_o or R_o , in place.

Formulas for the network elements are obtained by equating the poles and residues of the network impedance function to the given poles and residues of the general term of the series. Since both poles and residues occur in conjugate complex pairs, and since equality of real and imaginary parts is involved, there are four equations, which are necessary and sufficient to determine the four constants, R , L , G , C , of the network. The formulas that are obtained by solving these equations are given beneath Figs. 1 and 2.

The values given for G_o and R_o in each case assume that all the terms of the series are of the type specified for that case.

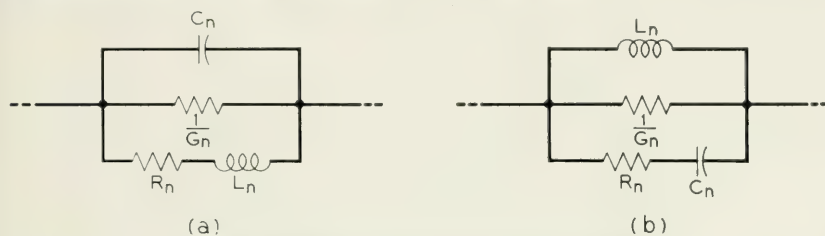


Fig. 2—General branches of the second kind.

Fig. 2a

(use where $F(p) = Z(p)$
and $Z_n(o) > 0$)

$$C_n = \frac{1}{2a_n}$$

$$\frac{1}{L_n C_n} = \beta_n^2 \left(\frac{b_n^2}{a_n^2} + 1 \right)$$

$$\frac{R_n}{L_n} = \frac{1}{a_n} (a_n \alpha_n - b_n \beta_n)$$

$$\frac{G_n}{C_n} = \frac{1}{a_n} (a_n \alpha_n + b_n \beta_n)$$

$$R_o = Z(0) - \sum_{n=\bullet}^{\infty} Z_n(0)$$

Fig. 2b

(use where $F(p) = Z(p)$
and $Z_n(o) < 0$)

$$C_n = \frac{\beta_n^2 (\alpha_n^2 + \beta_n^2)^2 (a_n^2 + b_n^2)}{2M^3}$$

$$\frac{1}{L_n C_n} = \frac{M^2}{\beta_n^2 (a_n^2 + b_n^2)}$$

$$R_n C_n = -\frac{a_n \alpha_n - b_n \beta_n}{M}$$

$$G_n L_n = \frac{N}{M(\alpha_n^2 + \beta_n^2)}$$

$$R_o = Z(0)$$

(7)

where $M = a_n(\beta_n^2 - \alpha_n^2) + 2\alpha_n \beta_n b_n$

$$N = -a_n \alpha_n^3 + 3\alpha_n^2 b_n \beta_n + 3a_n \alpha_n \beta_n^2 - b_n \beta_n^3$$

In the case of the parallel-type networks (Figs. 1a and 1b), $p_n = -\alpha_n + i\beta_n$ is a pole of the admittance, $Y(p)$, and $A_n = a_n + ib_n$ is the corresponding residue. In the case of the series-type network, the same symbols represent a pole and residue of the impedance, $Z(p)$.

The networks specified by Figs. 2a and 2b are duals of the networks of Figs. 1a and 1b, respectively, and are obtained from the latter merely by replacing L_n by C_n , R_n by G_n , and vice versa.

The formulas are intended to apply to complex poles. They can be applied to real poles by taking b_n and β_n equal to zero and doubling the residue, a_n , but this procedure is unnecessary, because the network rep-

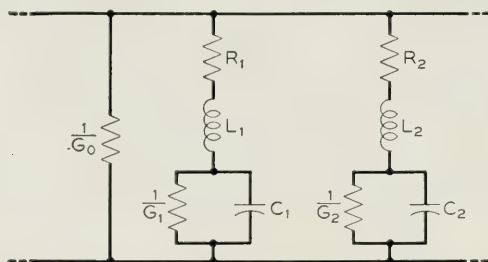


Fig. 3--Network of the first kind (branches 1a).

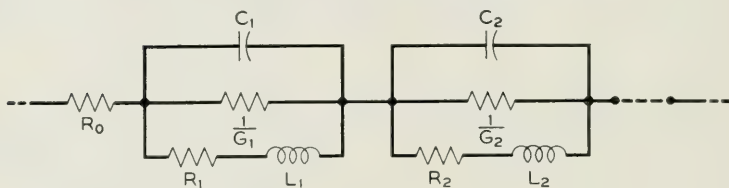


Fig. 4--Network of the second kind (branches 2a).

resentation of a real pole can be found readily enough by inspection of the impedance terms involved. (See Example 1.)

The above discussion is intended to sketch a general picture of the procedure. Individual cases may involve considerable detail that can be understood more readily by reference to the next section.

APPLICATIONS

Example 1a: A transmission line with its far terminals short-circuited affords a simple illustration of the equivalent network theory. Let it be assumed that the parameters, R , L , G and C of the line are constants. In the more advanced examples to follow, the variation of these parameters with frequency for a particular kind of line will be taken into consideration.

The impedance of the short-circuited line (Fig. 5) is

$$Z = Z_0 \tanh \Gamma \quad (1-0)$$

where Z_0 is the characteristic impedance and Γ is the total propagation constant of the line. We have

$$Z_0 = \left(\frac{R + pL}{G + pC} \right)^{1/2} \quad (1-1)$$

$$\Gamma = [(R + pL)(G + pC)]^{1/2} \quad (1-2)$$

R , L , G and C being given for the *total length* of line.

To obtain a development in terms of network branches of the kind shown in Fig. 1, we consider the admittance function,

$$Y = Y_0 \coth \Gamma \quad (1-3)$$

where $Y = 1/Z$ and $Y_0 = 1/Z_0$. Our first task is to find the poles of this function and the residues. Since the complex frequency variable p occurs

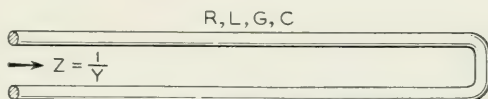


Fig. 5—Short-circuited transmission line.

under square roots in both Z_0 and Γ , it might be suspected, offhand, that the singularities of the function are branch points rather than poles. Such is not the case, however. There are no branch points and all the poles are simple.

The singularities of Y are to be found among the zeros of $\tanh \Gamma$, which occur at

$$\Gamma = i\pi n, \quad n = 0, \pm 1, \pm 2, \pm 3, \dots \quad (1-4)$$

To determine them, we solve

$$\Gamma^2 = (R + pL)(G + pC) = -\pi^2 n^2 \quad (1-5)$$

and find these roots:

$$p_n = -\alpha_n + i\beta_n, \quad p_{-n} = \bar{p}_n = -\alpha_n - i\beta_n$$

where

$$\alpha_n = \frac{G}{2C} + \frac{R}{2L}$$

$$\beta_n = \left[\frac{\pi^2 n^2}{LC} - \left(\frac{G}{2C} - \frac{R}{2L} \right)^2 \right]^{1/2} \quad (n > 0) \quad (1-6)$$

For $n = 0$, the above would give

$$p_0 = -\frac{R}{L}, -\frac{G}{C}$$

But if we let $\Gamma \rightarrow 0$, so that $\tanh \Gamma \rightarrow \Gamma$, we find that only the point, $-R/L$, is a singularity of Y ; the other point, $-G/C$, is a regular point. Therefore Y has only one real singularity.

To find the nature of the singularities of Y , we next calculate

$$\lim_{p \rightarrow p_n} \left[\frac{p - p_n}{Z_0(p) \tanh \Gamma(p)} \right] = A_n \quad (1-7)$$

and find that at each p_n the limit exists and has the value

$$A_n = \frac{1}{Z_0(p_n) \Gamma'(p_n)} = \frac{1}{L} - \frac{i \left(\alpha_n - \frac{R}{L} \right)}{\beta_n L} = a_n + ib_n \quad (1-8)$$

where $\Gamma'(p_n) = \frac{d}{dp} \Gamma(p)$, evaluated at $p = p_n$. The fact that this limit exists shows that all the singularities are simple poles. The values of A_n are then the residues at these poles.

When we now apply formulas (6) to determine the elements in the general branch of the equivalent network of Fig. 1a, we obtain, for $n > 1$,

$$L_n = \frac{L}{2}, \quad \frac{1}{L_n C_n} = \frac{\pi^2 n^2}{LC}, \quad \frac{G_n}{C_n} = \frac{G}{C}, \quad \frac{R_n}{L_n} = \frac{R}{L}. \quad (1-9)$$

The network then comprises an infinite number of such branches in parallel. Each branch has the same elements R_n and L_n , equal, respectively, to half the total resistance and inductance of the transmission line, but the elements G_n and C_n decrease from one branch to the next in inverse proportion to the squares of the integers.

The Q of the n^{th} branch, which can be regarded as the Q of the associated resonance of the short-circuited line, is

$$Q_n = \frac{\omega_n}{2\alpha_n} = \frac{\omega_n}{\frac{G_n}{C_n} + \frac{R_n}{L_n}} = \frac{\omega_n}{\frac{G}{C} + \frac{R}{L}} \quad (1-10)$$

where

$$\omega_n = \sqrt{\frac{1}{L_n C_n} - \frac{G_n^2}{C_n^2}} = \sqrt{\frac{\pi^2 n^2}{LC} - \frac{G^2}{C^2}} \quad (1-11)$$

Thus, for small dissipation, the resonances would become sharper in direct proportion to the frequency (if the parameters R , L , G , C , were invariable with frequency, as assumed).

The above described branches of the equivalent network account only for the complex poles ($n > 1$) of the admittance function. Two more branches remain to be calculated. One is for the real pole ($n = 0$), which occurs at $p_0 = -R/L$, with residue, $A_0 = \frac{1}{L}$. The required branch for this pole is

$$\frac{A_0}{p - p_0} = \frac{1}{R + pL} \quad (1-12)$$

The other is the final conductance branch, which is calculated as follows:

$$G_0 = Y(0) + \sum_{n=-\infty}^{\infty} \frac{A_n}{p_n} = \sqrt{\frac{G}{R}} \coth \sqrt{GR} - \frac{1}{R} - 2G \sum_{n=-\infty}^{\infty} \frac{1}{\pi^2 n^2 + GR} = 0 \quad (1-13)$$

so that, for this example, the conductance branch vanishes. The network is drawn in Fig. 6.

A series type of network, as shown in Fig. 7, can be determined by

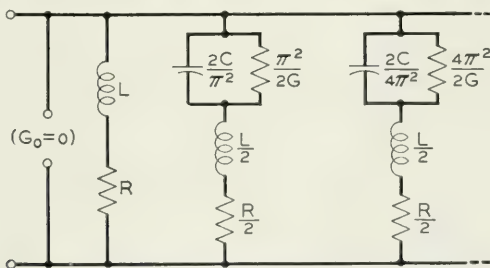


Fig. 6—Network of the first kind equivalent to the short-circuited line of Fig. 5.

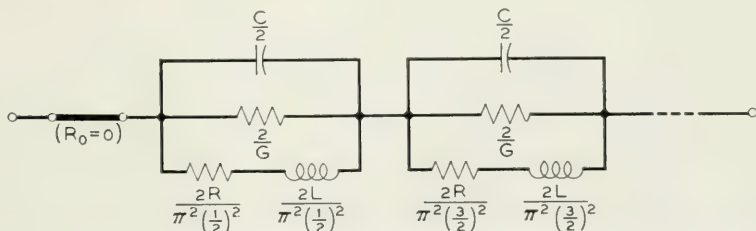


Fig. 7—Network of the second kind equivalent to the short-circuited line of Fig. 5.

similar means. Since, however, it is a dual of the parallel network of Fig. 9 for the open-circuited line, next to be discussed, it can be drawn immediately, without further calculation, once the latter has been found.

Example 1b: We now calculate a network for the same line with its far terminals open (Fig. 8). To obtain a network of the first kind, with branches in parallel, we deal with the admittance function,

$$Y = Y_0 \tanh \Gamma \quad (1-14)$$

The singularities of Y are found among the zeros of $\coth \Gamma$, which occur at

$$\Gamma = i\pi(n + \tfrac{1}{2}), \quad n = 0, \pm 1, \pm 2, \pm 3, \dots \quad (1-15)$$

The points $p = -R/L$ and $-G/C$ are *both* regular points this time. ($-G/C$ is a zero of Y .) The singularities are simple poles, as before, with residues,

$$A_n = \frac{1}{Z_0(p_n)\Gamma'(p_n)} \quad (1-16)$$

as before.

The network branches for the complex poles are therefore obtained merely by putting $n + \frac{1}{2}$ in place of the n in all formulas of the short-circuit network. There is no branch corresponding to the branch $R + pL$ of the other network and the conductance branch is again found to be zero. The complete parallel network is drawn in Fig. 9 and the series network, in Fig. 10.

It will be observed that the series network of Fig. 10 is the dual of the

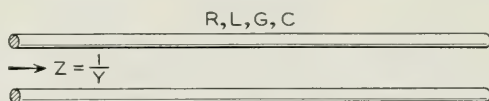


Fig. 8—Open-circuited transmission line.

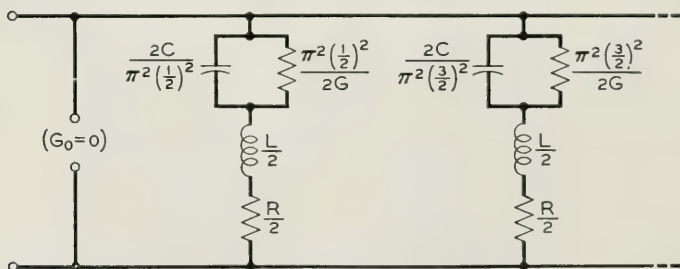


Fig. 9—Network of the first kind equivalent to the open-circuited line of Fig. 8.

parallel network of Fig. 6 and the series network of Fig. 7 is the dual of the parallel network of Fig. 9. These dual relationships are of course a result of the fact that the impedance of an open-circuited line is the dual of the impedance of the same line when short-circuited.

Example 2: Short-circuited Concentric Line (or Toroidal Cavity with E Radial). The preceding example considered a fictitious transmission line of invariable parameters, R, L, G, C , having a perfect short circuit at one end. The present example has to do essentially with the same problem but considers it from a more practical point of view. The variation of R and L with frequency is taken into account and the impedance of the "short-circuit" is no longer neglected.

Let the line be the piece of coaxial cable plugged at both ends with conducting material as illustrated in Fig. 11. Considered from an alternative point of view, our line is now a toroidal cavity oscillating in the

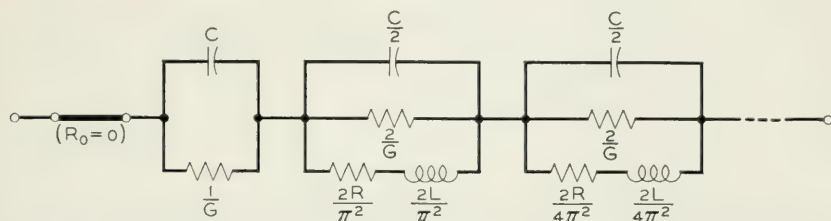


Fig. 10—Network of the second kind equivalent to the open-circuited line of Fig. 8.

mode where the electric force E is directed radially and the magnetic force H lies in planes at right angles to the axis. If we assume the cavity to be excited, or "driven," from one end,* the impedance that is effective in defining the selective characteristic of the cavity with respect to frequency is the total impedance at that end, that is, the sum of the impedance Z_1 , viewed into the cavity, and the impedance, Z_2 , of the adjacent end-plug. Therefore, we have to deal with the impedance,

$$Z = Z_1 + Z_2. \quad (2-1)$$

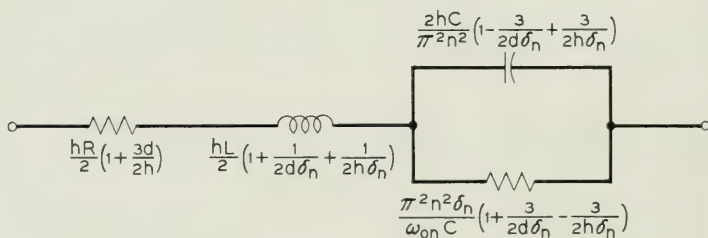
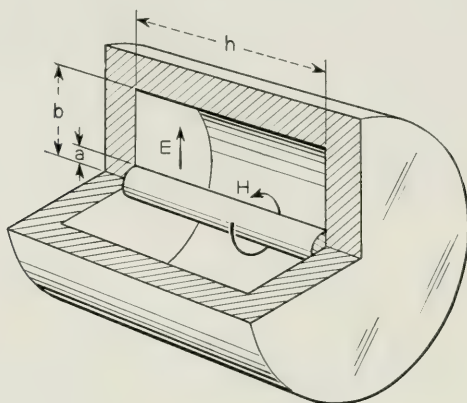
By "impedance" is here meant the same thing that one considers in looking at the problem from the point of view of transmission line theory, namely, the complex ratio, for exponential oscillations, of the voltage between the inside and outside cylindrical surfaces to the total current

* For determining the "natural frequencies" of oscillation of the cavity, it is immaterial at what point along it the impedance is taken; the total impedance at every point has the same roots. The impedance is, nevertheless, not the same at all points so that the behavior of the cavity, when driven, will depend to some extent on the driving point.

flowing axially in the inner conductor at the same point. The zeros of Z define the natural frequencies of oscillation of the cavity and their associated damping constants, or Q 's. Our task is to develop an equivalent network for this Z .

We have

$$Z = Z_1 + Z_2 = Z_0 \left(\frac{1 + \rho e^{-2\gamma h}}{1 - \rho e^{-2\gamma h}} + \frac{1 + \rho}{1 - \rho} \right) \quad (2-2)$$



$$R = \frac{1}{2\pi} \sqrt{\frac{\omega_{on} \mu}{2g}} \left(\frac{1}{a} + \frac{1}{b} \right)$$

$$\omega_{on} = \frac{\pi n v}{h}$$

$$L = \frac{\mu_0}{2\pi} \log \frac{b}{a}$$

$$v = \frac{1}{\sqrt{\mu_0 \epsilon_0}} = 3(10^8)$$

$$C = \frac{2\pi \epsilon_0}{\log \frac{b}{a}}$$

$$\delta_n = \sqrt{\frac{\omega_{on} \mu g}{2}}$$

(e.g.) $\mu_0 = 4\pi (10^{-7})$, $\epsilon_0 = \frac{10^{-9}}{36\pi}$, FOR AIR IN M.K.S. UNITS

$\mu = \mu_0$, $g = 5.8(10^7)$, FOR COPPER IN M.K.S. UNITS

($n = 1$ FOR FUNDAMENTAL MODE)

Fig. 11—Toroidal cavity, E radial.

where

$$Z_0 = \left(\frac{R + pL}{G + pC} \right)^{1/2}, \quad \gamma = (R + pL)^{1/2} (G + pC)^{1/2} \quad (2-3)$$

$$\rho = \frac{Z_2 - Z_0}{Z_2 + Z_0} \quad (2-4)$$

$$Z_2 = \frac{\eta}{2\pi} \log \frac{b}{a} \quad (2-5)$$

$$R + pL = \frac{\eta}{2\pi a} \frac{I_0(\sigma a)}{I_1(\sigma a)} + \frac{\eta}{2\pi b} \frac{K_0(\sigma b)}{K_1(\sigma b)} + \frac{p\mu_0}{2\pi} \log \frac{b}{a} \quad (2-6)$$

$$G + pC = \frac{2\pi}{\log \frac{b}{a}} (g_0 + p\epsilon_0) \quad (2-7)$$

h, a, b = cavity length, inner radius, outer radius, as shown in Fig. 11, all measured in meters

$$\eta = \left(\frac{p\mu}{g} \right)^{1/2} \quad (2-8)$$

$$\sigma = (p\mu g)^{1/2}$$

μ, g are permeability, conductivity of the conducting material of the walls (for copper: $\mu = 4\pi(10^{-7})$, $g = 5.8(10^7)$ in M.K.S. units).

μ_0, g_0, ϵ_0 are permeability, conductivity, dielectric constant of the dielectric material occupying the cavity (for air: $\mu_0 = 4\pi(10^{-7})$, $g_0 = 0$, $\epsilon_0 = (10^{-9})/36\pi$ in M.K.S. units), p = generalized frequency variable.

$I_0(z), I_1(z)$ are Bessel functions of the first kind for imaginary argument and of order 0, 1.

$K_0(z), K_1(z)$ are Bessel functions of the second kind for imaginary argument and of order 0, 1.

Except for ignored small deviations of the field around the corners of the cavity, the above formulas are exact. To arrive at results that are sufficiently compact to be useful, we make these approximations, at the start:

$$Z_0 = K_0 = \left[\frac{L}{C} \right]^{1/2} = \frac{\eta_0}{2\pi} \log \frac{b}{a}, \quad (2-9)$$

where

$$\eta_0 = \left[\frac{\mu_0}{\epsilon_0} \right]^{1/2} = 120\pi \text{ ohms} \quad (2-10)$$

From this,

$$\rho = \frac{\eta - \eta_0}{\eta + \eta_0} \quad (2-11)$$

Having in mind microwave applications, where the moduli of the arguments of the Bessel functions are >3000 , we take

$$\frac{I_0(z)}{I_1(z)} = \frac{K_0(z)}{K_1(z)} = 1$$

so that

$$R + pL = \frac{\eta}{2\pi} \left(\frac{1}{a} + \frac{1}{b} \right) + \frac{p\mu_0}{2\pi} \log \frac{b}{a} \quad (2-12)$$

Also, we have in mind only air dielectric and assume any loss therein to be negligible; that is, we assume $G = 0$.

All further approximations that are made are either

$$\frac{1}{1 - \Delta} \doteq 1 + \Delta \quad \text{or} \quad (1 + 2\Delta)^{1/2} \doteq 1 + \Delta$$

where, for an air-space enclosed by copper walls, and for frequencies on the order of 30,000 megacycles, Δ is on the order of 10^{-4} . For cavities made of other materials, the results obtained may not be sufficiently accurate and the problem would have to be reviewed from the start. In particular, the results do not hold for a cavity having walls of magnetic material, because we assume here that the permeability of the metal walls is the same as that of air; i.e., $\mu = \mu_0$.

To obtain an equivalent network of the first kind, we deal with the admittance, which is, from (2-2),

$$Y = \frac{1}{Z} = H_0 \frac{(1 - \rho)(1 - \rho e^{-2\gamma h})}{2(1 - \rho^2 e^{-2\gamma h})} \quad (2-13)$$

where $H_0 = 1/K_0$.

The poles of Y are then the zeros of $1 - \rho^2 e^{-2\gamma h}$, which are obtained by successive approximations. We first make a close estimate of the zeros by assuming that the impedance of the short-circuiting plugs is zero; that is, we assume, $Z_2 = 0$, whence $\rho = -1$. To obtain this estimate, we have to solve

$$\gamma h = \frac{p h}{v} \left(1 + \frac{2}{d\sigma} \right)^{1/2} = \pi i n \quad (n = \pm 1, \pm 2, \pm 3 \dots) \quad (2-14)$$

where

$$d = \frac{2ab \log(b/a)}{a + b}$$

and $v = 3(10^8)$ meters per second. The approximate solution is

$$p_{1n} = p_{0n} \left(1 + \frac{1}{d\sigma_{0n}} \right)$$

where

$$p_{0n} = \frac{i\pi nv}{h} \quad \text{and} \quad \sigma_{0n} = (p_{0n}\mu g)^{1/2}$$

Next we improve our estimate of the zeros by the well-known method involving the derivative of the function, $1 - \rho^2 e^{-2\gamma h}$, with respect to p , evaluated at p_{1n} . This now takes account of the actual impedance of the end-plugs. The values of the zeros, so obtained, are

$$p_n = -\alpha_n + i\beta_n, \quad p_{-n} = \bar{p}_n = -\alpha_n - i\beta_n$$

where

$$\begin{aligned} \alpha_n &= \omega_{0n} \left(\frac{1}{2d\delta_n} + \frac{1}{h\delta_n} \right) \\ \beta_n &= \omega_{0n} \left(1 + \frac{1}{2d\delta_n} - \frac{1}{h\delta_n} \right) \end{aligned} \tag{2-15}$$

where δ_n^* is the real part of σ_{0n} . That is,

$$\delta_n = (\omega_{0n}\mu g/2)^{1/2}$$

where

$$\omega_{0n} = \frac{\pi nv}{h}.$$

As an incidental matter of interest, the above gives the Q of the cavity at any resonance, namely

$$Q_n = \frac{\beta_n}{2\alpha_n} = d\delta_n \frac{1}{1 + \frac{2d}{h}} \tag{2-16}$$

For example, the dimensions, $a = .5$ cm., $b = 1.0$ cm., $h = .5$ cm. provide a cavity that resonates at about 30,000 megacycles. Then the Q 's at the first three resonances would be as follows:

n	$\frac{\omega_{0n}}{2\pi}$	Q
1	$30,000 \times 10^6$	4250
2	$60,000 \times 10^6$	6010
3	$90,000 \times 10^6$	7360

* For any frequency, $\delta = (\omega\mu g/2)^{1/2}$ is sometimes referred to as the "skin depth" because it is the depth of metal at which the current density falls to $1/e$ times its value at the surface of the metal.

The importance of including the effect of the end-plugs in determining Q is shown by the fact that, if they were assumed to have zero impedance, Q at the first resonance would be 12,120 instead of 4250.

To determine the residues at the poles, we write

$$Y = H_0 \frac{(1 - \rho)(1 - \rho e^{-2\gamma h})}{2(1 - \rho^2 e^{-2\gamma h})} = \frac{F(p)}{G(p)} \quad (2-17)$$

and then the residue at a simple pole p_n is

$$A_n = \frac{F(p_n)}{G'(p_n)} \quad (2-18)$$

This limit is found to exist, showing that the poles are, in fact, simple. The value found for the residue, A_n , is

$$\begin{aligned} A_n &= a_n + ib_n, \quad A_{-n} = \bar{A}_n = a_n - ib_n \\ a_n &= \frac{H_0 \omega_{0n}}{\pi n} \left(1 - \frac{1}{2d\delta_n} - \frac{1}{2h\delta_n} \right) \\ b_n &= \frac{H_0 \omega_{0n}}{\pi n} \left(\frac{1}{2d\delta_n} + \frac{1}{2h\delta_n} \right) \end{aligned} \quad (2-19)$$

When formulas (6) are applied to determine the elements of the tuned branches of the equivalent network of the first kind, the results are, for the n^{th} branch,

$$\begin{aligned} L_n &= \frac{K_0 \pi n}{2\omega_{0n}} \left(1 + \frac{1}{2d\delta_n} + \frac{1}{2h\delta_n} \right) \\ \frac{1}{L_n C_n} &= \omega_{0n}^2 \left(1 + \frac{1}{d\delta_n} - \frac{2}{h\delta_n} \right) \\ \frac{G_n}{C_n} &= \frac{\omega_{0n}}{2h\delta_n} \\ \frac{R_n}{L_n} &= \omega_{0n} \left(\frac{1}{d\delta_n} + \frac{3}{2h\delta_n} \right) \end{aligned} \quad (2-20)$$

In terms of the R , L and C of the piece of coaxial line, the elements of the n^{th} branch are as follows:

$$\begin{aligned} L_n &= \frac{hL}{2} \left(1 + \frac{1}{2d\delta_n} + \frac{1}{2h\delta_n} \right) \\ R_n &= \frac{hR}{2} \left(1 + \frac{3d}{2h} \right) \\ C_n &= \frac{2hC}{\pi^2 n^2} \left(1 - \frac{3}{2d\delta_n} + \frac{3}{2h\delta_n} \right) \\ G_n &= \frac{\omega_{0n} C}{\pi^2 n^2 \delta_n} \left(1 - \frac{3}{2d\delta_n} + \frac{3}{2h\delta_n} \right) \end{aligned} \quad (2-21)$$

where

$$L = \frac{\mu_0}{2\pi} \log \frac{b}{a}, \quad R = \frac{1}{2\pi} \left(\frac{\omega_{0n} \mu}{2g} \right)^{1/2} \left(\frac{1}{a} + \frac{1}{b} \right)$$

$$C = \frac{2\pi \epsilon_0}{\log \frac{b}{a}}, \quad \omega_{0n} = \frac{\pi n v}{h}, \quad \delta_n = \left(\frac{\omega_{0n} n \mu g}{2} \right)^{1/2}$$

The network is shown in Fig. 11.

It will be found that a "leakage" element, G_n , appears in the equivalent network, although the air dielectric in the cavity was assumed to have no leakage ($G = 0$). This element arises from the end-plugs and is necessary to account for the dissipation in them.

To obtain a network exactly equivalent to the cavity at all frequencies, we should add a branch corresponding to $n = 0$, as was done in example 1. This branch would make the equivalence hold down to and including zero frequency. But, inasmuch as the approximations that have been made hold only for the high frequencies, where the resonances occur, it would be inconsistent to add this branch. What has been arrived at, then, is a partial network representation that gives a close approximation to the impedance of the cavity at high frequencies, only.

Example 3: Toroidal Cavity with E Axial. For further illustration, we consider another mode of oscillation of the short-circuited concentric transmission line investigated in the previous example. This time it is assumed that the radial electric force vanishes while the axial electric force between the end-plugs exists. The magnetic force is directed in circles concentric with the cylindrical central conductor, as before. This situation is illustrated in Fig. 12, which is the same as Fig. 11, except for the new disposition of the E -vector.

For the new mode of oscillation, where the wave is a cylindrical one propagated back and forth between the inner and outer conducting cylinders, the oscillatory space is naturally thought of as a "toroidal cavity," while, in the previous example, where the wave was propagated axially back and forth between the terminal discs, the space was called a "concentric line." Actually, the cavity itself has the same geometric form in the two cases. A practical distinction may exist, however, in that the axial mode of oscillation could be more easily excited in a cavity whose axial length is large compared to its radius, while the cylindrical mode would arise more easily in a flat "pillbox" cavity whose radius is large compared to its axial dimension.

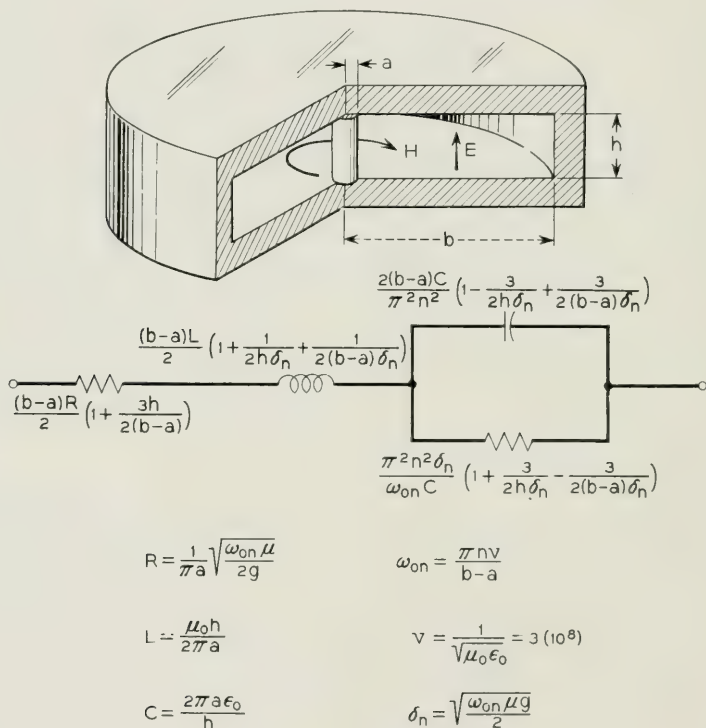
The approach to the problem will be that of transmission line theory, as before. This time, the "line" comprises two circular discs between

which the cylindrical wave is propagated. The series impedance and shunt admittance of such a line are functions of the radius and so will be designated $Z(r)$ and $Y(r)$, respectively. Their values are given below:

$$Z(r) = \frac{2\eta + i\omega\mu h}{2\pi r} \quad (3-1)$$

$$Y(r) = \frac{i\omega 2\pi r \epsilon_0}{h} \quad (3-2)$$

These formulas take into account the losses in the flat walls but assume the conductance of the air between them to be negligible. Losses in the inner and outer "short-circuiting" cylinders will be taken into account by the boundary conditions.



SEE FIGURE 11 FOR $\mu_0, \epsilon_0, \mu, g$

($n=1$ FOR FUNDAMENTAL MODE)

Fig. 12—Toroidal cavity, E axial.

If V is the voltage between the flat faces of the cavity at a radius r and I the total current in the lower face at this radius, we have

$$\begin{aligned}\frac{dV}{dr} &= -IZ(r) \\ \frac{dI}{dr} &= -VY(r)\end{aligned}\tag{3-3}$$

By differentiating,

$$\frac{d^2V}{dr^2} = -I \frac{dZ}{dr} - Z \frac{dI}{dr} = \left(\frac{1}{Z} \frac{dZ}{dr} \right) \frac{dV}{dr} + VZY\tag{3-4}$$

But

$$ZY = (2\eta + i\omega\mu h) \frac{i\omega\epsilon_0}{h} = \gamma^2$$

which is a squared propagation constant, *independent of r* , and

$$\frac{1}{Z} \frac{dZ}{dr} = -\frac{1}{r}$$

Therefore,

$$\frac{d^2V}{dr^2} + \frac{1}{r} \frac{dV}{dr} - \gamma^2 V = 0\tag{3-5}$$

is the differential equation for the voltage. The usual solution of this equation is a linear combination of $I_0(\gamma r)$ and $K_0(\gamma r)$ but since, in this case, the arguments will be almost purely imaginary, it is more convenient to employ the pair of functions, $J_0(-i\gamma r)$ and $N_0(-i\gamma r)$.

The solution for the voltage between the upper and lower surfaces at radius r is

$$V(r) = AJ_0(-i\gamma r) + BN_0(-i\gamma r)\tag{3-6}$$

and, from this, the total radial current in the lower surface, at that radius, is

$$I(r) = -\frac{1}{Z} \frac{dV(r)}{dr} = -iY_0(r)[AJ_1(-i\gamma r) + BN_1(-i\gamma r)]\tag{3-7}$$

where

$$Y_0(r) = 1/Z_0(r) = [Y(r)/Z(r)]^{1/2}$$

The impedance at the inner radius a , looking outward, is then

$$Z_1(a) = \frac{V(a)}{I(a)} = iZ_0(a) \frac{AJ_0(-i\gamma a) + BN_0(-i\gamma a)}{AJ_1(-i\gamma a) + BN_1(-i\gamma a)}\tag{3-8}$$

The total impedance at a (inward + outward) for which we require an equivalent network is

$$Z = Z_1(a) + Z_a$$

where Z_a is the impedance of the central plug to axial current, viz.,

$$Z_a = \frac{\eta h}{2\pi a} \frac{I_0(\sigma a)}{I_1(\sigma a)} \quad (3-9)$$

To evaluate the constants A and B , the following boundary conditions are imposed at radii a and b :

$$\text{at } a: V = V(a), \text{ a given voltage}$$

$$\text{at } b: V = I(b)Z_b$$

where Z_b is the impedance of the other "short circuit," comprising the outer cylindrical wall. It is given by

$$Z_b = \frac{\eta h}{2\pi b} \frac{K_0(\sigma b)}{K_1(\sigma b)} \quad (3-10)$$

Except for ignored small deviations of the field around the corners of the cavity, the above expressions are exact. The process of finding the singularities of Z by successive approximations results in expressions that are too long to write down here. To obtain results sufficiently compact for engineering use, we resort to the following asymptotic approximations for the Bessel functions:

$$\begin{aligned} J_0(z) &\sim \left(\frac{2}{\pi z}\right)^{1/2} \cos(z - \pi/4) \\ J_1(z) &\sim \left(\frac{2}{\pi z}\right)^{1/2} \cos(z - 3\pi/4) \\ N_0(z) &\sim \left(\frac{2}{\pi z}\right)^{1/2} \sin(z - \pi/4) \\ N_1(z) &\sim \left(\frac{2}{\pi z}\right)^{1/2} \sin(z - 3\pi/4) \\ \frac{I_0(z)}{I_1(z)} &\sim 1, \quad \frac{K_0(z)}{K_1(z)} \sim 1 \end{aligned} \quad (3-11)$$

Also, with an error on the order of 10^{-4} ,

$$Z_0(r) \sim \frac{h\eta_0}{2\pi r} = K_0(r) = 1/H_0(r)$$

These substitutions result in the following asymptotic formula for the total impedance Z at radius a

$$Z = K_0(a) \frac{\frac{2\eta}{\eta_0} \cos kx + i \left(1 + \frac{\eta^2}{\eta_0^2}\right) \sin kx}{\cos kx + \frac{i\eta}{\eta_0} \sin kx} \quad (3-12)$$

where $k = \frac{b}{a} - 1$ and $x = -i\gamma a$.

To find an equivalent network of the first kind to represent Z , we deal with the admittance, $Y = 1/Z$. It is instructive and saves much work to put Y in the form of exponential functions, with the substitution

$$\rho = \frac{\eta - \eta_0}{\eta + \eta_0}$$

which is the reflection coefficient at both inside and outside cylindrical surfaces of the cavity. By this means we obtain

$$Y = H_0(a) \frac{(1 - \rho)(1 - \rho e^{-2ikx})}{2(1 - \rho^2 e^{-2ikx})} \quad (3-13)$$

This is now identical in form to the formula (2-13) of example 2, where the E -vector was radially, instead of axially, directed. In fact, since

$$ikx = \gamma(b - a)$$

and

$$\gamma = \frac{i\omega}{v} \left(1 + \frac{2}{h\sigma}\right)^{1/2}$$

comparison with the similar formulas of example 2 shows that all the results of that example can be made to apply to the present one merely by changing the dimensional parameters as follows:

Example 2 (<i>E radial</i>)		Example 3 (<i>E axial</i>)
h	goes into	$b - a$
$d = \frac{2ab \log(b/a)}{a + b}$	goes into	h

The first result of interest is the value of Q , which is

$$Q_n = h\delta_n \frac{1}{1 + \frac{2h}{b - a}} \quad (3-14)$$

where, as before, δ_n is the "skin depth" equal to the real part of σ_n . That is,

$$\delta_n = \sqrt{\frac{\omega_{0n} \mu g}{2}}$$

To gain an idea of numerical magnitudes, consider the same cavity used in example 2. The dimensions are, as before, $h = .5$ cm., $b - a = .5$ cm. For the square cross-section chosen, the first resonance again occurs at 30,000 megacycles, very nearly, and we can make the following direct comparison of the Q 's for the two modes of oscillation:

n	$\omega_{0n}/2\pi$	Q_n	
		Ex. 2 (E radial)	Ex. 3 (E axial)
1	$30,000 \times 10^6$	4250	4370
2	$60,000 \times 10^6$	6010	6180
3	$90,000 \times 10^6$	7360	7560

Due to the asymptotic approximations used, the results for example 3 are not as accurate as those for example 2; the two sets of results show only that the Q of the cavity is substantially the same for the two different modes of oscillation.

The poles of Y are given by

$$\begin{aligned}
 p_n &= -\alpha_n + i\beta_n, & p_{-n} = \bar{p}_n &= -\alpha_n - i\beta_n \\
 \alpha_n &= \omega_{0n} \left[\frac{1}{2h\delta_n} + \frac{1}{(b-a)\delta_n} \right] \\
 \beta_n &= \omega_{0n} \left[1 + \frac{1}{2h\delta_n} - \frac{1}{(b-a)\delta_n} \right]
 \end{aligned} \tag{3-15}$$

and the residues are

$$\begin{aligned}
 A_n &= a_n + ib_n, & A_{-n} = \bar{A}_n &= a_n - ib_n \\
 a_n &= \frac{H_0(a)\omega_{0n}}{\pi n} \left[1 - \frac{1}{2h\delta_n} - \frac{1}{2(b-a)\delta_n} \right] \\
 b_n &= \frac{H_0(a)\omega_{0n}}{\pi n} \left[\frac{1}{2h\delta_n} + \frac{1}{2(b-a)\delta_n} \right]
 \end{aligned} \tag{3-16}$$

Applying formulas (6) gives the following values for the n^{th} branch of

the network of the first kind:

$$\begin{aligned}
 L_n &= K_0(a) \frac{\pi n}{2\omega_{0n}} \left[1 + \frac{1}{2h\delta_n} + \frac{1}{2(b-a)\delta_n} \right] \\
 \frac{1}{L_n C_n} &= \omega_{0n}^2 \left[1 + \frac{1}{h\delta_n} - \frac{2}{(b-a)\delta_n} \right] \\
 \frac{G_n}{C_n} &= \frac{\omega_{0n}}{2(b-a)\delta_n} \\
 \frac{R_n}{L_n} &= \omega_{0n} \left[\frac{1}{h\delta_n} + \frac{3}{2(b-a)\delta_n} \right]
 \end{aligned} \tag{3-17}$$

in all of which $\omega_{0n} = \pi n v / (b - a)$ and $v = 1/(\mu_0 \epsilon_0)^{1/2} = 3(10^8)$ meters per second.

The results can be put in the same form as those obtained for the other cavity mode, dealt with in example 2, by employing the "primary constants" of the cylindrical transmission line, viz.:

$$\begin{aligned}
 R(a) &= \frac{1}{\pi a} \left[\frac{\omega_{0n} \mu}{2g} \right]^{1/2} & L(a) &= \frac{\mu h}{2\pi a} \\
 G(a) &= 0 & C(a) &= \frac{2\pi a \epsilon_0}{h}
 \end{aligned}$$

In terms of these constants, the elements of the n^{th} branch of the equivalent network of the first kind are

$$\begin{aligned}
 L_n &= \frac{(b-a)L(a)}{2} \left(1 + \frac{1}{2h\delta_n} + \frac{1}{2(b-a)\delta_n} \right) \\
 R_n &= \frac{(b-a)R(a)}{2} \left(1 + \frac{3h}{2(b-a)} \right) \\
 C_n &= \frac{2(b-a)C(a)}{\pi^2 n^2} \left(1 - \frac{3}{2h\delta_n} + \frac{3}{2(b-a)\delta_n} \right) \\
 G_n &= \frac{\omega_{0n}C(a)}{\pi^2 n^2 \delta_n} \left(1 - \frac{3}{2h\delta_n} + \frac{3}{2(b-a)\delta_n} \right)
 \end{aligned} \tag{13-8}$$

The network is shown in Fig. 12.

As in the preceding example, a leakage element arises, in spite of the fact that we assumed initially that g_0 of the air in the cavity is zero. This element accounts for the losses in the inner and outer cylindrical walls.

A number of people with whom the above material has been discussed have given helpful comments and criticisms. I wish to acknowledge my debt in this respect to H. Nyquist, S. A. Schelkunoff, R. M. Foster, S. O. Rice, J. Riordan and W. H. Wise.

BIBLIOGRAPHY

1. R. M. Foster, "A Reactance Theorem," *Bell System Tech. J.*, 1924, **3**, p. 259; also, "Theorems on the Driving Point Impedance of Two-Mesh Circuits," *Bell System Tech. J.*, 1924, **3**, p. 651.
2. W. Cauer, "Die Verwirklichung von Wechselstromwiderständen vorgeschriebener Frequenzabhängigkeit," *Archiv für Electrot.*, 1926-27, **17**, p. 355.
3. G. A. Campbell and R. M. Foster, "Fourier Integrals for Practical Applications," *Bell System Tech. J.*, Oct. 1928, pp. 639-707; *Bell System Monograph* B-584.
4. T. C. Fry, "The Use of Continued Fractions in the Design of Electric Networks," *Bull. of Am. Math. Soc.*, 1929, **35**, p. 463.
5. O. Brune, "Synthesis of a Finite Two-terminal Network Whose Driving-point Impedance Is a Prescribed Function of Frequency," *Journal of Math. and Physics (M.I.T.)*, Oct. 1931, **1**, p. 191.
6. Sidney Darlington, *Journal of Math. and Physics (M.I.T.)*, Sept. 1939, **4**, pp. 257-353.
7. E. C. Titchmarsh, "The Theory of Functions," Oxford Univ. Press, 2nd ed., 1939, p. 110.
8. Gustav Doetsch, *Theorie und Anwendung der Laplacetransformation*, 1st Am. Ed. 1943, p. 139.
9. S. A. Schelkunoff, "Representation of Impedance Functions in Terms of Resonant Frequencies," *Proc. of the I.R.E.*, 1944, **32**, No. 2, p. 83.
10. H. W. Bode, *Network Analysis and Feedback Amplifier Design*, D. Van Nostrand Co., 1945.
11. P. I. Richards, "A Special Class of Functions with Positive Real Part in a Half-plane," *Duke Math. J.*, 1947, **14**, pp. 777-786. Also "General Impedance Function Theory," *Quart. Appl. Math.*, 1948, **6**, pp. 21-29.
12. E. A. Guillemin, *The Mathematics of Circuit Analysis*, John Wiley and Sons, Inc., New York, 1950, pp. 409-422.

Abstracts of Bell System Technical Papers* Not Published in This Journal

Universal Equalizer Chart. D. A. ALSBERG¹. *Electronics*, **24**, pp. 132, 134, Nov., 1951.

Modification of familiar Smith chart consolidates on one time-saving plot all positive-value solutions to the two general equations for series, shunt, and bridged-T audio equalizers.

Limits on the Energy of the Antiferromagnetic Ground State. P. W. ANDERSON¹. *Phys. Rev.*, **83**, p. 1260, Sept. 15, 1951.

Post-War Achievements of Bell Laboratories, I. O. E. BUCKLEY¹. *Bell Tel. Mag.*, **30**, pp. 163-173, Autumn, 1951.

Filamentary Growths on Metal Surfaces—"Whiskers". K. G. COMPTON¹, A. MENDIZZA¹, and S. M. ARNOLD¹. *Corrosion*, **7**, pp. 327-334, Oct., 1951. (Monograph 1885).

Filamentary growths have been found on metal surfaces of some of the parts used in telephone communications equipment, particularly on parts shielded from free circulation of air. The growths are of the same character as those known as "whiskers," which developed between the leaves of cadmium plated variable air condensers and caused considerable trouble in military equipment during the early part of World War II. An investigation has been under way in an attempt to determine the mechanism of growth of the whiskers, found not only on cadmium plated parts but also on other metals. This paper summarizes the findings to date as revealed by the study of approximately one thousand test specimens of different metals, solid and plated, exposed under various environmental conditions. The study is being extended in the light of the findings which have developed during the course of the work.

An Unattended Broad-Band Microwave Repeater for the TD-2 Radio Relay System. R. W. FRIIS¹ and K. D. SMITH¹. *Elec. Eng.*, **70**, pp. 976-981, Nov., 1951. (Similar article in *The Bell System Technical Journal*,

* Certain of these papers are available as Bell System Monographs and may be obtained on request to the Publication Department, Bell Telephone Laboratories, Inc., 463 West Street, New York 14, N. Y. For papers available in this form, the monograph number is given in parentheses following the date of publication, and this number should be given in all requests.

¹ B. T. L.

October, 1951, Part II, entitled *The DT-2 Microwave Radio System* reprinted as Monograph 1921).

To meet the stringent requirements of the 4,000-mile transcontinental microwave relay system, a number of new developments had to be included in the design of the repeater stations. The circuits of these unattended stations, and how they are maintained, are the subject of this article.

The Bell System's Part in Defending the Nation. F. R. KAPPEL². *Bell Tel. Mag.*, **30**, pp. 141-152, Autumn, 1951.

Quickly and accurately checks performance of private or common-carrier p-m or f-m mobile telephone transmitters, such as those used in 30 to 44 -mc highway and 152 to 175-mc urban service. Measures r-f power output, audio sensitivity, signal-to-noise ratio and harmonic distortion and gives speech intelligibility check in few minutes.

Mobile Transmitter Testing Set. G. J. KENT³. *Electronics*, **24**, pp. 106-109, Nov., 1951.

A New Electrolysis Switch for Underground Lead Sheath Cable Drainage Systems. V. B. PIKE¹. *Corrosion*, **7**, p. 1, Oct., 1951.

A High Temperature Stage for the Polarizing Microscope. E. A. WOOD¹. *Am. Mineral.*, **36**, pp. 768-772, Sept.-Oct., 1951.

A Precise Sweep-Frequency Method of Vector Impedance Measurement. D. A. ALSBERG¹. *Proc. I.R.E.*, **39**, pp. 1393-1400, Nov., 1951. (Monograph 1911).

The impedance of a two-terminal network is defined completely by the insertion loss and phase shift it produces when inserted between known sending and receiving impedances. Recent advances in precise wide-band phase and transmission measuring circuits have permitted practical use of this principle. Reactive and resistive impedance components are read directly from a simple graphical chart in which frequency is not a parameter. The basic principle described promises attractive possibilities in many cases of impedance measurements where present methods are inadequate.

Electron-Vibration Interactions and Superconductivity. J. BARDEEN¹. *Revs. Modern Phys.*, **23**, pp. 261-270, July, 1951. (Monograph 1912).

The Copper Oxide Rectifier. W. H. BRATTAIN¹. *Revs. Modern Phys.*, **23**, pp. 203-212, July, 1951.

It is shown that the conductivity in the ohmic part of the cuprous oxide layer can be explained with the usual band picture of semiconductors only by assum-

² A. T. & T. Co.

³ W. E. Co.

ing the presence of some donor-type impurities in addition to the usual acceptor type. The energy difference between the acceptors and the filled band is 0.3 electron volt, and the total number of impurity atoms is about 10^{14} to 10^{16} per cm^3 , the number of donors being less than but of the same order as the number of acceptors. One finds that the density of ion charge in the rectifying layer is of the same order of magnitude as the difference between the donors and acceptors found from the conductivity. The field at the copper-cuprous oxide interface is about 2×10^4 volts/cm; the height of the potential at the surface as compared with the oxide interior is about 0.5 volt; and the thickness of the space charge layer about 5.0×10^{-5} cm. The diffusion equation for flow of current through this space charge region can be integrated to give the current in terms of the field at the interface and the applied potential across the space charge layer. Two currents are involved, one from the semi-conductor to the metal (I_s) and one from the metal to the semiconductor (I_m) which is similar to a thermionic emission current into the semiconductor. The net current is, of course, $I = I_m - I_s$. One can get this "emission" current (I_m) by dividing the true current by the factor $1 - \exp(-eV_a/kT)$, where V_a is the applied potential. This emission current depends on the absolute temperature and on the field at the copper-cuprous oxide interface. At high fields the logarithm of the current is proportional to the square root of the field, and at low fields the current decreases more rapidly indicating a patchy surface having small areas of low potential maximum from which all the emission comes when the field is large.

Effect of Packaging on Corrosion of Zinc Plated Equipment. K. G. COMPTON¹, S. M. ARNOLD¹, and A. MENDIZZA¹. *Corrosion*, **7**, pp. 365-372, Nov., 1951.

Physics as a Science and an Art. K. K. DARROW¹. *Phys. Today*, **4**, pp. 6-11, Nov., 1951. (Monograph 1914).

The last of six invited papers presented on October 25th during the symposium on "physics today" which keynoted the 20th Anniversary Meeting of the American Institute of Physics in Chicago.

Ionization by Electron Impact in CO, N₂, NO, and O₂. H. D. HAGSTRUM¹. *Revs. Modern Phys.*, **23**, pp. 185-203, July, 1951. (Monograph 1916).

Ionization by electron impact in diatomic gases has been studied in this work with a mass spectrometer designed to measure m/e , appearance potential, and initial kinetic energy for each ion observed. Results have been obtained for the gases CO, N₂, NO, and O₂ with some confirmatory work in H₂. Discussion is included of the nature and identification of dissociative ionization processes and of the retarding potential and appearance potential measurements. Values of important quantities such as the dissociation energies of CO, N₂, and NO; the sublimation energy of C; the electron affinity of O; and the excitation energy of O⁻ are determined again by electron impact in this work.

¹ B. T. L.

Equivalent Temperature of an Electron Beam. M. E. HINES¹, Letter to the Editor., *J. Appl. Phys.*, **22**, pp. 1385-1386, Nov., 1951.

Bell System Cable Sheath Problems and Designs. F. W. HORN¹ and R. B. RAMSEY¹, *Elec. Eng.*, **70**, pp. 1070-1075, Dec., 1951. (Monograph 1917).

Engineering Planning. H. S. OSBORNE², *J. Eng. Educ.*, **42**, pp. 121-125, Nov., 1951.

Acceptance Inspection of Purchased Material. J. E. PALMER³ and E. G. D. PATERSON¹, *Ind. Quality Control*, **8**, pp. 23-27, Nov., 1951.

A Note on the Partial Differential Equations Describing Steady Current Flow in Intrinsic Semiconductors. R. C. PRIM¹, Letter to the Editor. *J. Appl. Phys.*, **22**, pp. 1388-1389, Nov., 1951.

General Theory of Symmetric Biconical Antennas. S. A. SCHELKUNOFF¹, *J. Appl. Phys.*, **22**, pp. 1330-1332, Nov., 1951. (Monograph 1922).

This paper presents the input admittance of a symmetric biconical antenna of an arbitrary angle as the limit of a certain sequence of functions. The first term of this sequence approaches the exact expressions for the input admittance as the cone angle approaches either zero or 90°. For this reason our conjecture is that this term represents a good first approximation for all angles.

Artificial Dielectrics for Microwaves. W. M. SHARPLESS¹, *Proc. I. R. E.*, **39**, pp. 1389-1393, Nov., 1951. (Monograph 1923).

This paper presents a procedure for measuring the dielectric properties of metal-loaded artificial dielectrics in the microwave region by the use of the short-circuited line method. Formulas, based on transmission-line theory, are included and serve as guides in predicting the approximate dielectric properties of certain loading configurations.

¹ B. T. L.

² W. E. Co.

Contributors to this Issue

W. O. BAKER, B.S., Washington College, Maryland, 1935; Ph.D., Princeton University, 1938; Bell Telephone Laboratories, 1939-. Dr. Baker has carried on investigations of the molecular structure and physical properties of polymers, particularly the fundamental constitution of synthetic rubbers and plastics. Harvard Fellowship, 1936-37 and Proctor Fellowship, 1938-39. Member of American Chemical Society, American Physical Society, and American Society for Testing Materials.

J. H. HEISS, B.S. in Ch.E., Newark College of Engineering, 1942; Bell Telephone Laboratories, 1934-. Mr. Heiss has devoted his time to studying experimental wire coating procedures and the test methods involved, the experimental production of high polymers (plastics) and the examination of their physical properties, and the properties of high polymers in solution. Member of American Chemical Society.

W. S. HAYWARD, JR., A.B., Harvard University, 1943; S.M., Harvard Graduate School of Engineering, 1947; Aircraft Radio Laboratory, June-December 1943; U. S. Navy, Aviation Electronics Officer, 1944-46; Bell Telephone Laboratories, 1947-. Mr. Hayward has taught telephone switching circuit design at the Laboratories and is currently making probability studies of telephone traffic.

BROCKWAY McMILLAN, B.S., Massachusetts Institute of Technology, 1936; Ph.D., Massachusetts Institute of Technology, 1939; Instructor of Mathematics, Massachusetts Institute of Technology, 1936-39; Proctor Fellow and Henry B. Fine Instructor in Mathematics, Princeton University, 1939-42; U.S.N.R., 1942-46, studying exterior ballistics of guns and rockets; Los Alamos Laboratory, Spring 1946; Bell Telephone Laboratories, 1946-. Dr. McMillan has been engaged in mathematical research and consultation work. Member of American Mathematical Society, Institute of Mathematical Statistics, and A.A.A.S.

R. E. STAEBLER, B.E.E., College of the City of New York, 1947; M.E.E., Polytechnic Institute of Brooklyn, 1948; U. S. Army 1943-46, Communications Officer; Instructor in Electrical Engineering, Polytechnic Institute of Brooklyn, Fall, 1950; Bell Telephone Laboratories, 1948-. After completing the Communications Development Training Program,

with rotational assignments in the transmission, switching, and apparatus departments, Mr. Staehler became a member of a switching group concerned with both local and toll signaling. He is working at present on a new voice-frequency toll signaling development. Member of Tau Beta Pi and Eta Kappa Nu.

M. K. ZINN, B.S. in E.E., Purdue University, 1918; U. S. Army, 1918-19. Amer. Tel. and Tel., 1919-1934; Bell Telephone Laboratories, 1934-. As a transmission engineer, Mr. Zinn has been concerned with land and submarine loaded cables, telephone instruments, buried cable and submarine cable with amplifiers and special problems. Member of A. I. E. E. and Tau Beta Pi.

Technical Journal

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

VOLUME XXXI

MAY 1952

NUMBER 3

Present Status of Transistor Development

J. A. MORTON 411

An Experimental Electronically Controlled Automatic Switching
System

W. A. MALTHANER AND H. EARLE VAUGHAN 443

New Techniques for Measuring Forces and Wear in Telephone
Switching Apparatus

WARREN P. MASON AND SAMUEL D. WHITE 469

A Comparison of Signalling Alphabets

E. N. GILBERT 504

Principal Strains in Cable Sheaths and Other Buckled Surfaces

I. L. HOPKINS 523

A New Recording Medium for Transcribed Message Services

JAMES Z. MENARD 530

Introduction to Formal Realizability Theory-II

BROCKWAY MCMILLAN 541

Abstracts of Bell System Technical Papers not Published in this
Journal

601

Contributors to This Issue

608

THE BELL SYSTEM TECHNICAL JOURNAL

PUBLISHED SIX TIMES A YEAR BY THE
AMERICAN TELEPHONE AND TELEGRAPH COMPANY

195 BROADWAY, NEW YORK 7, N. Y.

CLEO F. CRAIG, *President*

CARROLL O. BICKELHAUPT, *Secretary*

DONALD R. BELCHER, *Treasurer*

EDITORIAL BOARD

F. R. KAPPEL

O. E. BUCKLEY

H. S. OSBORNE

M. J. KELLY

J. J. PILLIOD

A. B. CLARK

R. BOWN

D. A. QUARLES

F. J. FEELY

PHILIP C. JONES, *Editor*

M. E. STRIEBY, *Managing Editor*

SUBSCRIPTIONS

Subscriptions are accepted at \$3.00 per year. Single copies are 75 cents each.

The foreign postage is 65 cents per year or 11 cents per copy.

Present Status of Transistor Development

By J. A. MORTON

(Manuscript received March 17, 1952)

The invention of the transistor provided a simple, apparently rugged device that could amplify—an ability in which the vacuum tube had long held a monopoly. As with most new electron devices, however, a number of extremely practical limitations had to be overcome before the transistor could be regarded as a practical circuit element. In particular: the reproducibility of units was poor—units intended to be alike were not interchangeable in circuits; the reliability was poor—in an uncomfortably large fraction of units made, the characteristics changed suddenly and inexplicably; and the “designability” was poor—it was difficult to make devices to the wide range of desirable characteristics needed in modern communications functions. This paper describes the progress that has been made in reducing these limitations and extending the range of performance and usefulness of transistors in communications systems. The conclusion is drawn that for some system functions, particularly those requiring extreme miniaturization in space and power as well as reliability with respect to life and ruggedness, transistors promise important advantages.

INTRODUCTION

When the transistor was announced not quite four years ago; it was felt that a new departure in communication techniques had come into view. Here was a mechanically simple device which could perform many of the amplification functions over which the electron tube had long held a near monopoly. The device was small, required no heater power, and was potentially very rugged; moreover, it consisted of materials which might be expected to last indefinitely long, and it did not appear to be too complicated to make.

However, as might be expected for a newly invented electron device, the practical realization of these promises still required the overcoming

of a number of obstacles. While the operation of the first devices was well understood in a general way, several items were limiting and puzzling, for example:

a—Units intended to be alike varied considerably from each other—the *reproducibility was bad*.

b—In an uncomfortably large fraction of the exploratory devices, the properties changed suddenly and inexplicably with time and temperature, whereas other units exhibited extremely stable characteristics with regard to time—the *reliability was poor*.

c—It was difficult to use the theory and then existing undeveloped technology to develop and design devices to a varied range of electrical characteristics needed for different circuit functions. Performance characteristics were limited with respect to gain, noise figure, frequency range and power—the *designability was poor*.

Before the transistor could be regarded as a practical circuit element, it was necessary to find out the causes of these limitations, to understand the theory and develop the technology further in order to produce and control more desirable characteristics.

Over the past two years measurable progress has been made in reducing, *but not eliminating*, the three listed limitations.

These advances have been obtained through an improved understanding, improved processes and very importantly through improved germanium materials. As a result:

a—the beginnings of method have evolved in the use of the theory to explain and predict the electrical network characteristics of transistors in terms of physical structure and material properties.

b—It is now possible to evaluate some of the effects and physical meaning of empirically derived processes and thereby to devise better methods subject to control. Previously, inhomogeneities in the material properties masked the dependence of the transistor electrical properties even on bulk properties (such as resistivity) as well as on processing effects.

c—As a result, on an exploratory development level, it is now possible to make transistors in the laboratory to several sets of prescribed characteristics with usable tolerances and satisfactory yields.

d—Such transistors are greatly improved over the old ones in so far as life and ruggedness are concerned, and some reduction in temperature dependence has been achieved. However, it is not to be inferred that all reliability problems are solved.

e—It has become possible in the laboratory to explore experimentally some of the consequences of the theory with the result that point con-

tact devices with new ranges of performance are indicated. Even more importantly, new p - n junction devices have been built in the laboratory and these junction devices have indicated an extension in several performance characteristics.

f—By having interchangeable and reliable devices with a wider range of characteristics, it has become possible to carry on exploratory circuit and system applications on a more realistic basis. Such applications effort is, in turn, stimulating the development of new devices towards new characteristics needed by these circuit and system studies.

It is the purpose of the remainder of this paper to give an over all but brief summary of recent progress made at Bell Telephone Laboratories in reducing the above-mentioned limitations on reproducibility, reliability and performance. Since a fair number of types of devices are currently under development, each with different characteristics to be optimized, the data will be presented as a sort of montage of characteristics of several different types of devices. It is not to be inferred that any one type of transistor combines all of the virtues any more than such a situation exists in the electron tube art. Moreover, it will be impossible in a paper of practical length to present complete detailed characteristics on all or even several of these devices under development; nor would it be appropriate since most of these data are on devices currently under development. Rather, what is desired, is a summary of progress across the board to give the reader an integrated and up-to-date picture of the current state of transistor electronics.

REPRODUCIBILITY STATUS

Description of Transistors

Before quantitative data comparing the characteristics of past with present transistors are presented, it will be useful to briefly review physical descriptions of the various types of transistors to be discussed. Fig. 1 shows a cutaway view of the now familiar point-contact cartridge type transistor. All of the early transistors were of this general construction and the characteristics of a particular one, called the Type A¹, will be used as a reference against which to measure results now obtainable with new types under current development. Fig. 2 is a semi-schematic picture of the physical operation of such a device. Pressing down upon the surface of a small die of n -type germanium are two rectifying metal electrodes, one labelled E for emitter, the other C for collector. A third electrode, the base, is a large area ohmic contact to the underside of the die of germanium. The emitter and collector electrodes obtain their

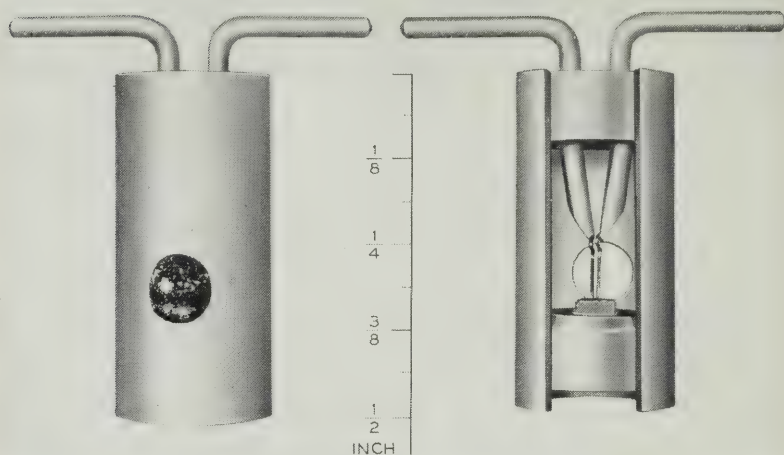


Fig. 1—The type A transistor structure.

rectifying properties as a result of the p - n barrier (indicated by the dotted lines) existing at the interface between the n -type bulk material and small p -type inserts under each point. When the collector is biased with a moderately large negative voltage (in the reverse direction) so that the collector barrier has relatively high impedance, a small amount of reverse current flows from the collector to the base in the form of electrons as indicated by the small black circles. Now, if the emitter is biased a few tenths of a volt positively in the forward direction, a current of holes (indicated by the small open circles) is injected from the

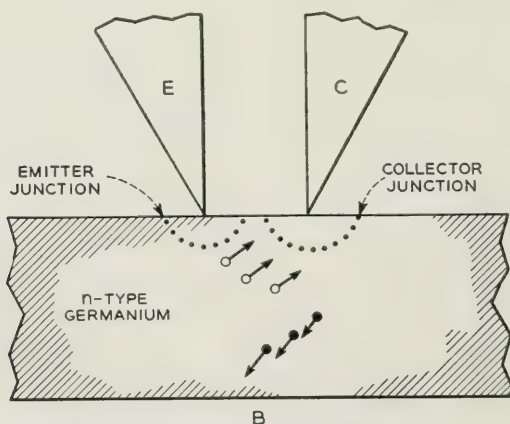


Fig. 2—Schematic diagram of a point-contact transistor.

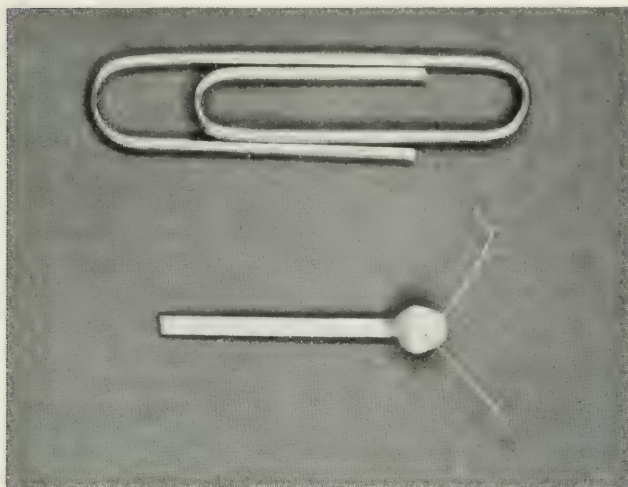


Fig. 3—The M1689 point-contact transistor is typical of those used in miniature packaged circuit functions.

emitter region into the n -type material. These holes are swept along to the collector under the influence of the field initially set up by the original collector electron current—thus adding a controlled increment of collector current. Because of their positive charge these holes can lower the potential barrier to electron flow from collector to base and thus allow several electrons to flow in the collector circuit for every hole entering the collector barrier region. This ratio of collector current change to emitter current change for fixed collector voltage is called alpha, the current gain. In point-contact transistors alpha may be larger than unity. Since the collector current flows through a high impedance when the emitter current is injected through a low impedance, voltage amplification is obtained as well.

Some of the new transistors are point-contact transistors similar in physical appearance to the type A. However, their electrical characteristics will be shown to be significantly improved not over the old type A only insofar as reproducibility and reliability are concerned, but also as to range of performance.

For use in miniature packaged circuit functions, the point contact transistor has been miniaturized to contain only its bare essentials. Fig. 3 is a photograph of a so-called "bead" transistor (compared to a paper clip for size) and several of the current development types are being made in this form.

In Fig. 4 is shown the family of static characteristics representative of the M1689 bead type transistor. Note in particular the collector

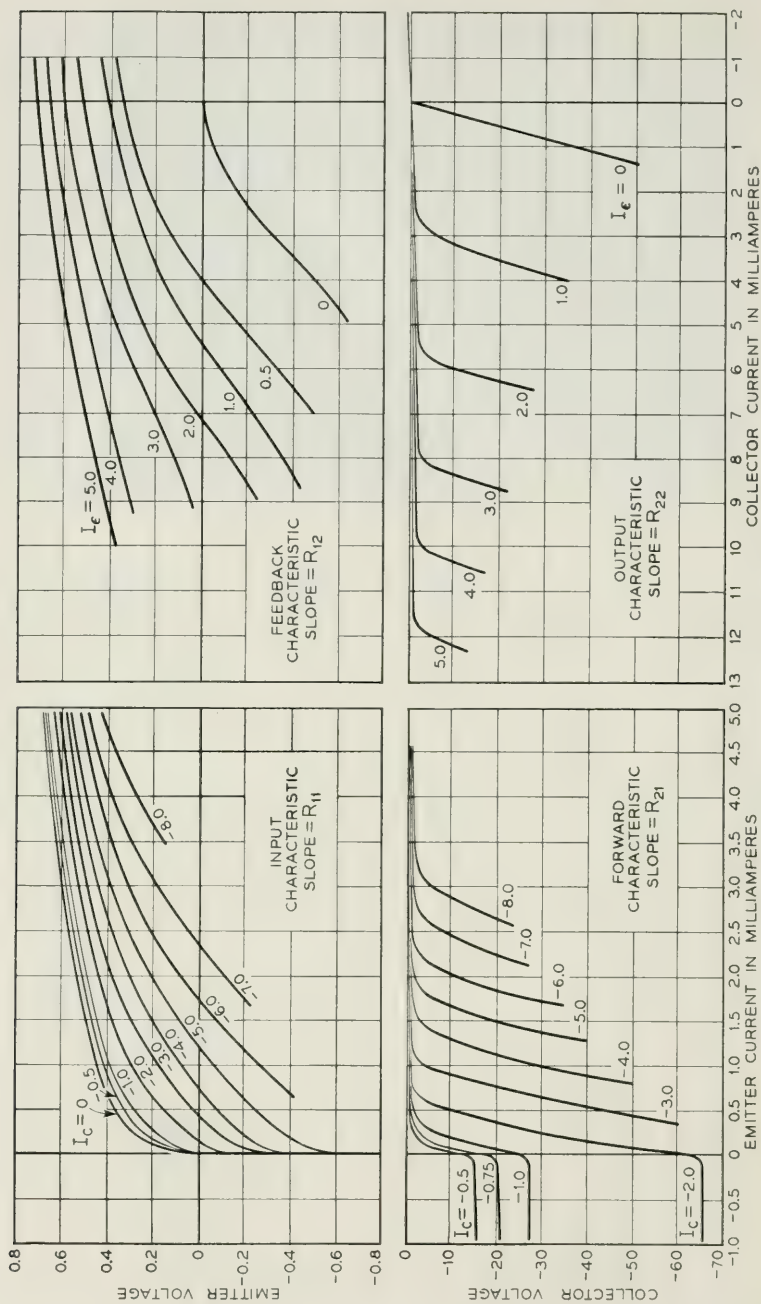


Fig. 4—Static characteristics of the M1689 transistor.

family which gives the dependence of collector voltage upon collector current with emitter current as parameter. These characteristics may be thought of as the dual to the plate family of a triode.² The slope of these curves is very nearly the small-signal ac collector impedance of the transistor.* For a fixed collector voltage of -20 volts, when the emitter current is changed from zero to one milliamperes, note that the collector current correspondingly changes slightly more than two milliamperes, indicating a current gain, α , of slightly more than two.

Newest member of the transistor family recently described by Shockley, Sparks, Teal, Wallace and Pietenpol is the n - p - n junction transistor.^{3, 4} Fig. 5 is a schematic diagram of such a structure. In the center of a bar of single crystal n -type germanium there is formed a thin layer

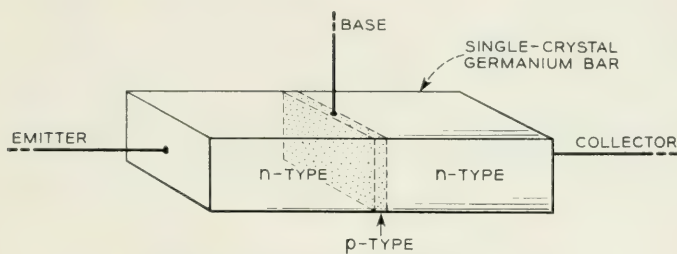


Fig. 5—The n - p - n junction transistor

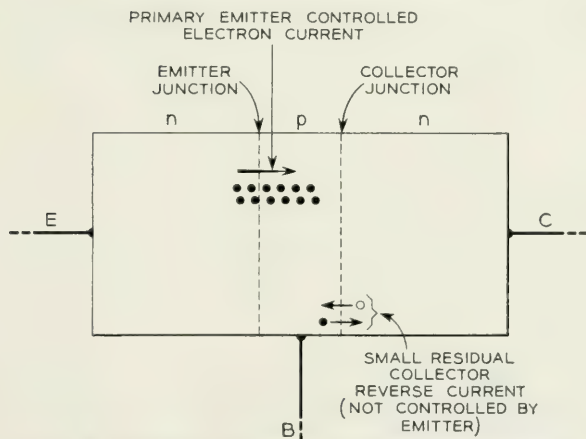


Fig. 6—Schematic diagram of a junction transistor.

* As shown by Ryder and Kircher,¹ the ac collector impedance, $r_c = R_{22} - R_{12}$, where R_{22} is the open-circuited output impedance and R_{12} is the open-circuit feedback impedance. Usually, $R_{22} \gg R_{12}$.

of p -type germanium as part of the same single crystal. Ohmic non-rectifying contacts are securely fastened to the three regions as shown, one being labelled emitter, one base and one collector. In many simple respects, except for change in conductivity type from p - n - p in the point-contact (see Fig. 2) to n - p - n in the junction type, the essential behavior is similar.

As shown in Fig. 6, if the collector junction is biased in the reverse direction, i.e., electrode C biased positively with respect to electrode B, only a small residual back current of holes and electrons will diffuse across the collector barrier as indicated. However, unlike the point-contact device, this reverse current will be very much smaller and relatively independent of the collector voltage because the reverse impedance of such bulk barriers is so many times higher than that of the barriers produced near the surface in point-contact transistors. Now again, if the emitter barrier is biased in the forward direction, a few tenths of a volt negative with respect to the base is adequate, then a relatively large forward current of electrons will diffuse from the electron-rich n -type emitter body across the reduced emitter barrier into the base region. If the base region is adequately thin so that the injected electrons do not recombine in the p -type base region (either in bulk or on the surface), practically all of the injected emitter current can diffuse to the collector barrier; there they are swept through the collector barrier field and collected as an increment of controlled collector current. Hence, again, since the electrons were injected through the low forward impedance and collected through the very high reverse impedance of bulk type p - n barriers, very high voltage amplification will result. No current gain is possible in such a simple bulk structure and the maximum attainable value of alpha is unity. However, because the bulk barriers are so much better rectifiers than the point surface barriers, the ratio of collector reverse impedance to emitter forward impedance is many times greater, more than enough to offset the point-contact higher alpha; thus, the junction unit may have much larger gain per stage.^{1, 3, 4} Fig. 7 is a photograph of a developmental model of such a junction transistor called the M1752.

The upper part of Fig. 8 is a collector family of static characteristics for the M1752 n - p - n junction transistor. By way of comparison to those of the point contact family, note the much higher reverse impedance of the collector barrier (relatively independent of collector voltage) and the correspondingly smaller collector currents when the emitter current is zero. In fact, Fig. 9 is an expanded plot of the lower left rectangle of the collector family of Fig. 8. The almost ideal straight-line character



Fig. 7—The M1752 junction transistor.

and regular spacing of these curves persists down to voltages as low as 0.1 volt and currents of a few microamperes. Thus, essentially linear Class A amplification is possible for as little collector power as a few microwatts. Constant collector power dissipation curves of 10, 50 and 100 microwatts are shown dotted for reference.

Reproducibility of Linear Characteristics

In describing progress in the reproducibility of those transistor characteristics pertinent to small-signal linear applications, one possible method is to give the statistical averages and dispersions in the linear open-circuit impedances of the transistor as defined by Messrs. Ryder and Kircher.¹ Such a procedure, of course, implies a state of statistical control in the processes leading to a reasonably well behaved normal distribution for which averages and control limits can be defined. This situation can be said to be in effect for most transistors under current development.

However, for the old type A unit, control simply was not in evidence; so that in quoting figures on type A's, ranges for commensurate fractions of the total family will be given. In order that symbols and terminology

will be clear, it will be useful to review briefly the method of defining the linear characteristics of all transistors. In Fig. 10 is shown a generalized network representing the transistor in which the input terminals are emitter-base and the output terminals are collector-base. Then, over a sufficiently small region of the static characteristics, the linear relations between the incremental emitter and collector voltages and currents may be represented by the pair of linear equations shown.¹

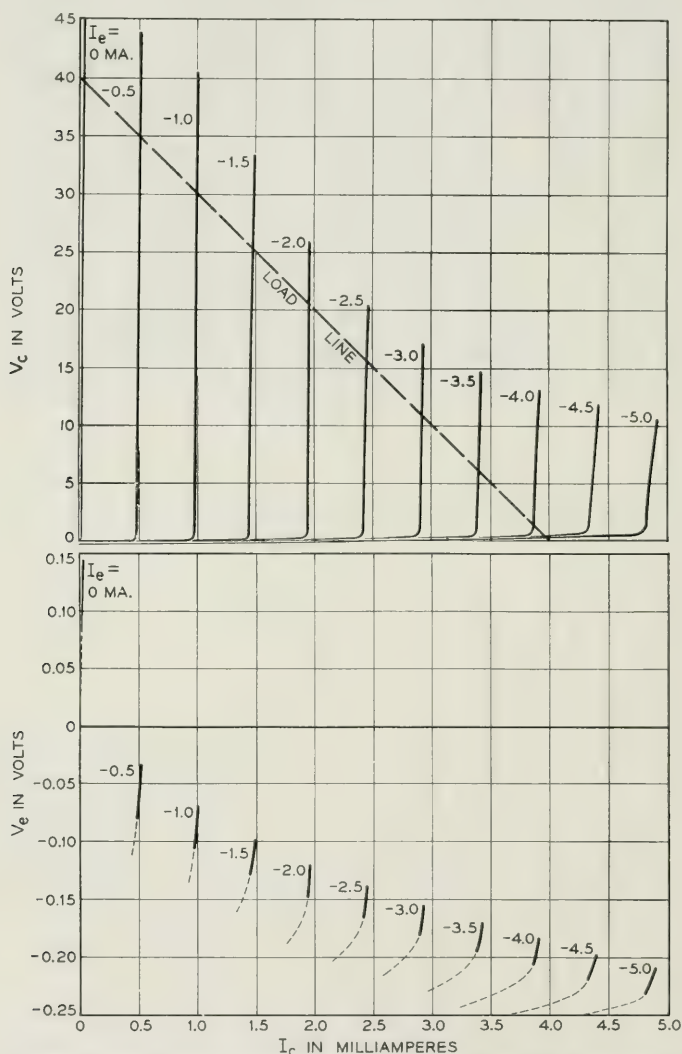


Fig. 8—Static characteristics of the M1752 junction transistor.

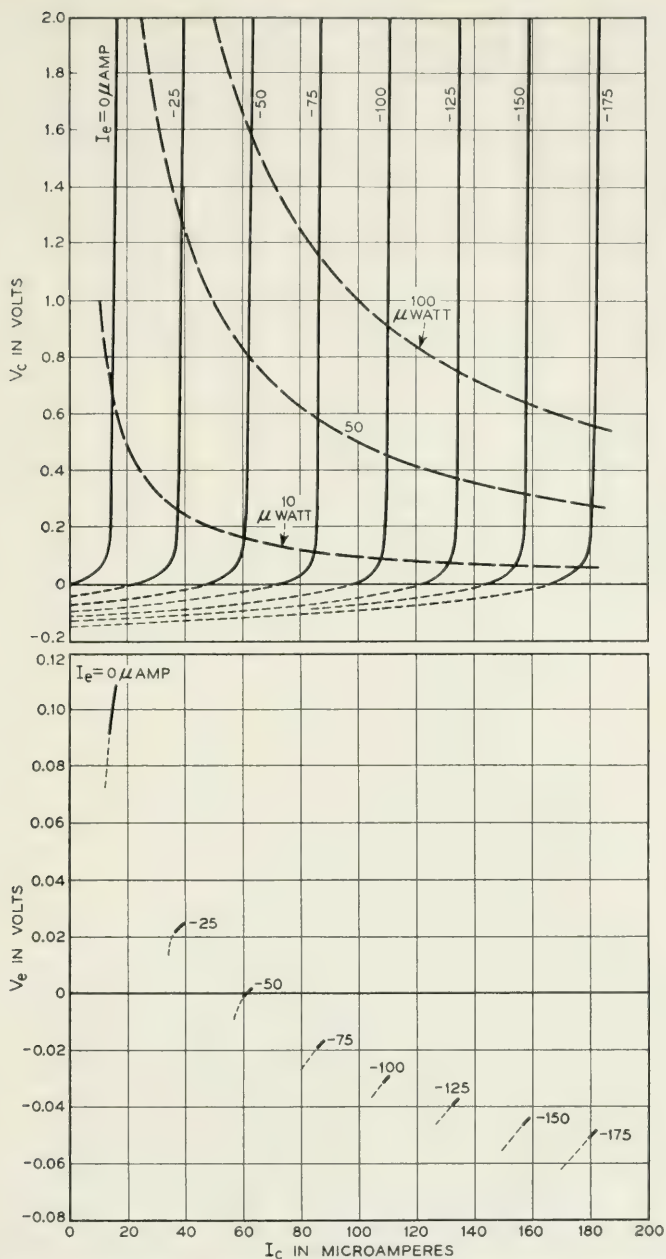


Fig. 9—Expanded plot of the microwatt region of the static characteristics of the M1752 transistor.

The coefficients are simply the open-circuit driving point and transfer impedances of the transistor, or the slopes of the appropriate static characteristics at fixed dc operating currents. These equations may be represented by any one of a large number of equivalent circuits of which the one shown in Fig. 11 is perhaps currently most useful. In this circuit r_e is very nearly the ac forward impedance of the emitter barrier, r_c is very nearly the ac reverse impedance of the collector barrier, r_b is the feedback impedance of the bulk germanium common to both, and a is the circuit current gain representing carrier collection and multiplication if any. It turns out this is very nearly equal to the current multiplication factor a of the collector barrier mentioned before. Average values of these elements for the type A transistor are given in Fig. 11. In Fig. 12 are given the ranges of these parameters for the type A as of September, 1949, and the control limits* for the same characteristics for new point-contact transistors now under development. For September, 1949, the ranges are taken about the average values shown in Fig. 11 for the type A transistor. The control limits given for the present situation apply to a number of different types of point contact transistors so that the present average values of these

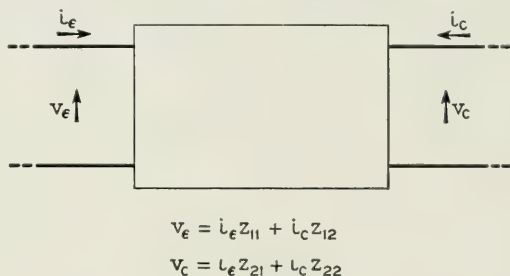


Fig. 10—The general linear transistor.

equivalent circuit elements depend upon the type of transistor considered. In Fig. 13 are given the average values of the characteristics of the M1729 point-contact video amplifier transistor which bears the closest resemblance to the older type A transistor. By way of contrast are given some typical values of the elements for the M1752 junction transistor which is not yet far enough along in its development to have design centers fixed nor reliable dispersion figures available.

As Ryder and Kircher have shown,¹ transistors in the grounded-base connection may be short-circuit unstable if $a > 1$ and r_b is too large,

* A.S.T.M. Manual, "Quality Control of Materials," Jan. 1951, Part III, pp. 55-114.

since r_b appears as a positive feedback element. The curve in Fig. 14 is a plot of the short-circuit stability contour when r_e and r_c have the nominal values of 700 and 20,000 ohms. Transistors having a and r_b sufficiently large to place their representative points above this contour will be short-circuit unstable, i.e., they will oscillate when short-circuited. Those having an $a - r_b$ point below the stability contour will be unconditionally stable under any termination conditions. The large unshaded rectangle bounds those values of a and r_b , which were repre-

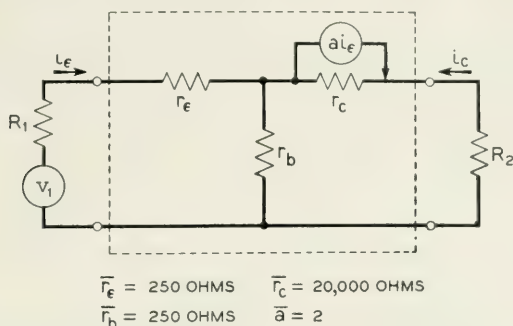


Fig. 11—Equivalent circuit and average element values of the type A transistor.

ELEMENT	RANGE SEPTEMBER 1949	RANGE JANUARY 1952
a	4 : 1	$\pm 20\%$
r_c	7 : 1	$\pm 30\%$
r_e	3 : 1	$\pm 20\%$
r_b	7 : 1	$\pm 25\%$

Fig. 12—Reproducibility of point-contact linear characteristics.

TYPE	M 1729	M 1752
r_e	120	25
r_b	75	250
r_c	15,000	5×10^6
a	2.5	0.95

Fig. 13—Average characteristics of the M1729 and typical characteristics of the M1752 transistors.

sentative of the type A transistor in September, 1949. It is apparent that the circuit user of type A units had approximately a 50 per cent chance of obtaining a short-circuit unstable unit from a large family of type A units. The smaller shaded rectangle bounds the values of a and r_b now realized in the M1729 transistor presently under development. Not only has the spread in characteristics been greatly reduced as shown, but also the design centers have been moved to a region for which all members of the M1729 family are unconditionally stable.

It is of interest to note that spreads of the order of ± 20 to ± 25 per cent are of the same magnitude as those dispersions now existing amongst the characteristics of presently available well-controlled electron tubes. These kinds of data on reproducibility of the linear equivalent circuit element values hold for practically all classes of point-contact devices

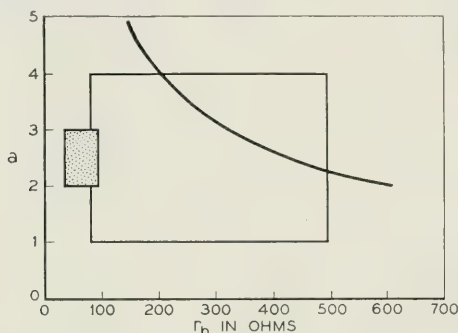


Fig. 14—Stability contour and ranges of a and r_b .

now under development for cw transmission service. While it is too early to prove that such a situation pertains as well to junction transistors, there is every reason to expect similar results after a suitable development period.

Reproducibility of Large-Signal Characteristics for Pulse Application

When electron devices are employed for large-signal applications, particularly those of switching and computing, it is well known that the characteristics must be controlled over a very broad range of variables from cutoff to saturation. In September, 1949, very little attempt was made to control such pulse use characteristics. In the intervening time, transistor circuit studies have proceeded to the point where it is possible to define certain necessary large scale transistor characteristics which, if met, permit such transistors to be used interchangeably and reproducibly in a variety of pulse circuit functions such as binary counters,

bit registers, regenerative pulse amplifiers, pulse delay amplifiers, gated amplifiers and pulse generators. Moreover, it has been possible to meet these requirements on a developmental level with good yields in at least three types of point-contact switching transistors. The scope of this paper will not permit a detailed accounting of the technical features of this situation and such an account will be forthcoming in future papers on these particular studies. However, a brief description of some of the more important pulse characteristics and their tolerances is certainly pertinent.

In practically all of the transistor pulse handling circuits examined to date, one characteristic common to all is the ability of the transistor, by virtue of its current gain, to present various types of two-state negative resistance characteristics at any one or all of its pairs of terminals. A typical simple circuit and corresponding characteristic is shown in Fig. 15 for the emitter-ground terminals when a sufficiently large value of resistance is inserted in the base to make the circuit unstable. In region I where the emitter is negative, the input resistance is essentially the reverse characteristic of the emitter as a simple diode. In region II as the emitter goes positive, alpha, the current gain rises rapidly above unity. If R_b is sufficiently large and alpha, the current gain, is greater

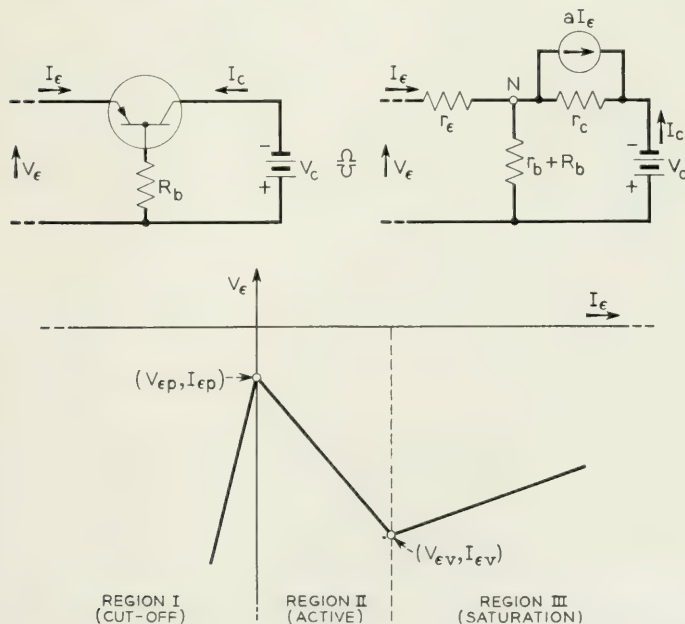


Fig. 15—Emitter-ground negative resistance circuit and characteristic.

than unity the emitter to ground voltage will begin to fall because of the larger collector current increments driving the voltage of the node N negative more rapidly than the emitter current drop through r_e would normally carry it. This transition point is called the peak point. If then $\alpha(r_b + R_b)$ is sufficiently large, in this sense, the input resistance may be negative in this region II. When the internal node voltage has fallen to a value near that of the collector terminal the "valley point" has been reached. At this point, the emitted hole current has reduced the collector impedance to a minimum value beyond which a is essentially zero; the transistor is said to be saturated. From this point on the input impedance again becomes positive and is determined almost entirely by the base and emitter impedances. By terminating the emitter-ground terminals in various ways with resistor-capacitor-bias combinations, such a network can be made to perform monostable, astable or bistable functions. Under such conditions, the emitter current and correspondingly the collector current switch back and forth between cutoff and saturation values. For example, in Fig. 16 is shown a value of emitter bias and load resistance such that there are three possible equilibrium values of emitter current and voltage. It may be shown that the two intersections in regions I and III are stable whereas that in region II is unstable. Hence, if the stable equilibrium is originally in I, a small positive pulse Δ_p applied to the emitter will be enough to switch from stable point I to stable point II and conversely, $-\Delta_e$ will carry it from the high current point to the low current point. The circuit designer is interested in reproducing in a given circuit (with different transistors of the same type) the following points of the characteristic:

a—The off impedance of the emitter—he desires that this be greater than a certain minimum.

b—The peak point V_{ep} —he desires that this be smaller than a certain maximum.

c—The value of the negative resistance—he desires that this be greater than a certain minimum.

d—The valley point V_{es} , I_{es} —he desires that these be greater than certain minima, and

e—The slope in region III—he desires that this be smaller than a certain maximum so that he may control it by external means.

It may be shown that these conditions can be satisfied for useful circuits by specifying certain maximum and minimum boundaries on the static characteristics. Fig. 17 is an idealized set of input or emitter characteristics. By specifying a minimum value for the reverse resistance

in region I, condition (a) above is satisfied. By specifying a maximum slope in region II and III, condition (e) is satisfied. Now refer to the idealized collector family in Fig. 18; by specifying a maximum value to V_{c3} , it is possible to insure condition (d) and by specifying a minimum value for r_{co} , condition (b) can be satisfied. Finally, in Fig. 19 by de-

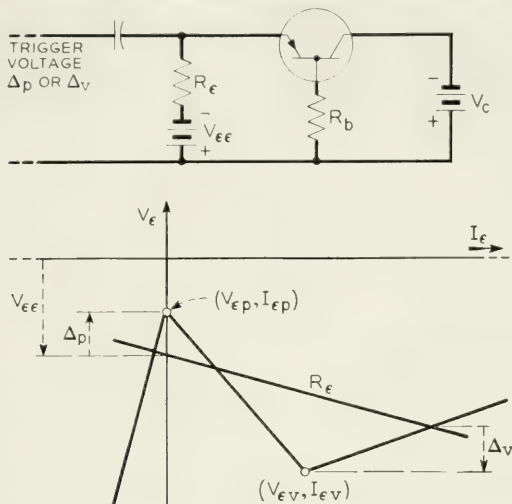


Fig. 16—Bistable circuit and characteristics showing trigger voltage requirements.

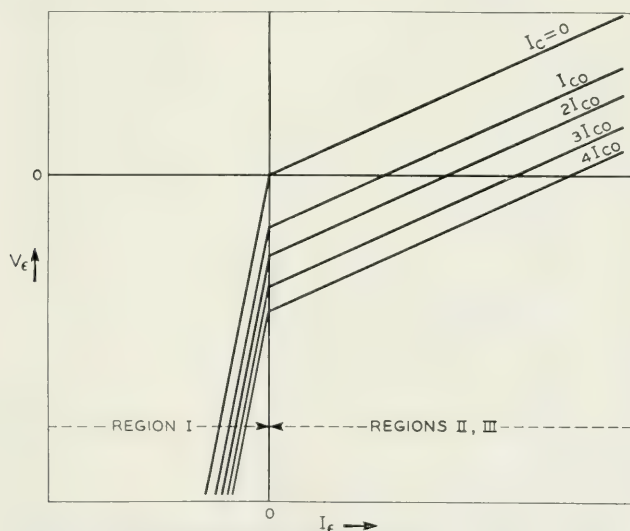


Fig. 17—Idealized emitter characteristics — slope = R_{11} .

manding that α , as a function of I_e , go through a transition from a negligible value (at small negative I_e) to a value well in excess of unity (at a correspondingly small positive value of I_e) and maintain its value well in excess of unity at large values of I_e , conditions (b) and (c) can be met.

In Fig. 20 are given the characteristic specifications which must be met by the M1689 bead type switching transistor now under development. With these kinds of limits, circuit users find it possible to interchange such M1689 units in various pulse circuits and obtain overall circuit behavior reproducible to the order of about ± 2 db.

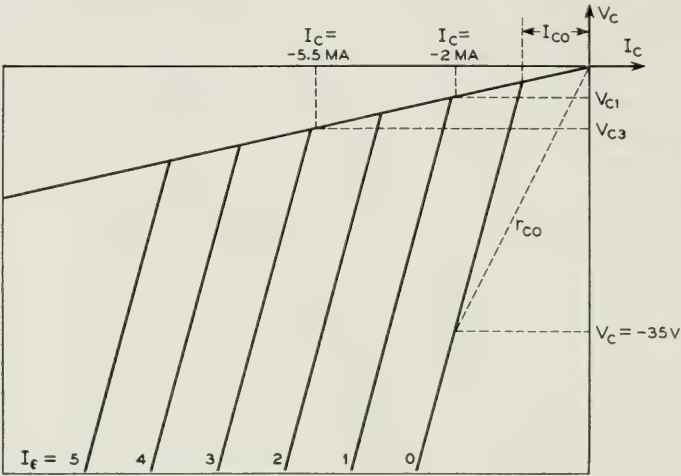


Fig. 18—Idealized collector characteristics.

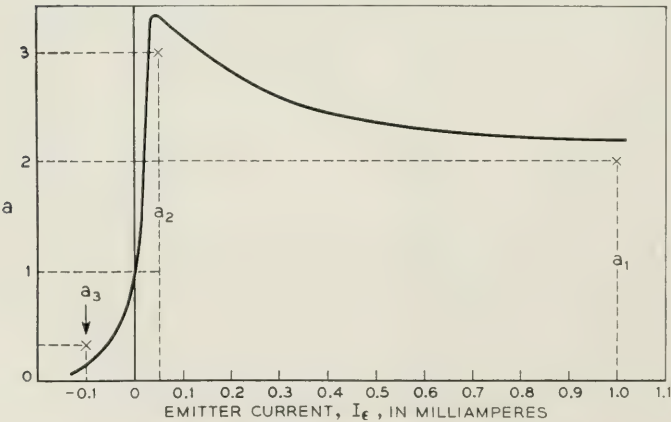


Fig. 19—Effective alpha characteristic.

TEST	CONDITIONS	MINIMUM	MAXIMUM
r_{c0} —OFF COLLECTOR DC RESISTANCE	$V_C = -35$ V DC $I_E = 0$ MA DC	17,500 OHMS	—
V_{C1} —ON COLLECTOR VOLTAGE	$I_C = -2$ MA DC $I_E = 1$ MA DC	—	-3V DC
V_{C3} —ON COLLECTOR VOLTAGE	$I_C = -5.5$ MA DC $I_E = 3$ MA DC	—	-4V DC
OFF EMITTER RESISTANCE	$V_C = -10$ V DC	50,000 OHMS	—
ON EMITTER RESISTANCE R_{11}	$V_C = -10$ V DC $I_E = 1$ MA DC	—	800 OHMS
a_1	$V_C = -30$ V DC $I_E = 1.0$ MA DC	1.5	—
a_2	$V_C = -30$ V DC $I_E = +0.05$ MA DC	2.0	—
a_3	$V_C = -30$ V DC $I_E = -0.1$ MA DC	—	0.3
R_{12} —OPEN CIRCUIT FEEDBACK RESISTANCE	$V_C = -10$ V DC $I_E = +1$ MA DC	—	500 OHMS
R_{21} —OPEN CIRCUIT FORWARD RESISTANCE	$V_C = -10$ V DC $I_E = +1$ MA DC	15,000 OHMS	—
R_{22} —OPEN CIRCUIT OUTPUT RESISTANCE	$V_C = -10$ V DC $I_E = +1$ MA DC	10,000 OHMS	—

Fig. 20—Tentative characteristics for the M1689 switching transistor.

RELIABILITY FIGURE OF MERIT	SEPTEMBER 1949	JANUARY 1952
AVERAGE LIFE	$\approx 10,000$ HOURS	$> 70,000$ HOURS
EQUIVALENT TEMPERATURE COEFFICIENT OF r_c	-1% PER DEG C	-1/4% PER DEG C
SHOCK	?	$> 20,000$ G
VIBRATION	?	20-5000 CPS NEGLECTIBLE TO 100 G

Fig. 21—Reliability status.

RELIABILITY STATUS

Life

Reliability figures of merit are not too well defined for electron tubes and the same situation certainly holds at present for transistors. However, insofar as these quantities can be presently defined, Fig. 21 shows

a comparison between the present status and that in September, 1949. Estimates of the half-life of a statistical family of devices are at best arbitrary and necessarily amount to extrapolations of survival curves assuming that a known survival law will continue to hold.* In September, 1949, life tests on type A units had been in effect some 4000 hours. With the assumption of an exponential survival law, it was not possible, on the basis of a 4000 hour test, to estimate the slope sufficiently accurately to warrant a half-life estimate in excess of 10,000 hours. These same type A units have now run on life test for approximately 20,000 hours. With the more reliable estimate of survival slope now possible, the half-life is now estimated to be somewhat in excess of 70,000 hours. It should be emphasized, however, that these are type A units of more than two years ago made with inferior materials and processes. It is believed that those units under current development, being made with new materials and processes, are superior; but, of course, life tests are only a few thousand hours old. Although these new data are encouraging, it is still too early to extrapolate the data such a long way.

Temperature Effects

Transistors like other semiconductor devices are more sensitive to temperature variations than electron tubes. In terms of the linear equivalent circuit elements, the collector impedance, r_c , and the current gain, a are the most sensitive. Over the range from -40°C to 80°C the other elements are relatively much less sensitive. For type A transistors these temperature variations in r_c and a are shown in Fig. 22. While these curves are definitely not linear, an average temperature coefficient for r_c of about -1 per cent per degree was estimated for the purpose of easy tabulation and comparison in Fig. 21.

Thus, for the early type A, r_c fell off to about 20 to 30 per cent of its room temperature value when the temperature was raised to $+80^{\circ}\text{C}$; at the same time a increased from 20 to 30 per cent over the same temperature range. Today, this variation has been reduced by a factor of about four for r_c in most point-contact types, the variations in the current gain being relatively unchanged. Fig. 23 illustrates the temperature dependence of r_c and a for the M1729 transistor now under development. Again, for purposes of easy comparison in Fig. 21, the actual dependence of Fig. 23 was approximated by a linear variation and

* Estimates of life, of course, depend upon definitions of "death". For these experiments, the transistors were operated as Class A amplifiers. A transistor is said to have failed when its Class A gain has fallen 3 db or more below its starting value.

only the slope given in Fig. 21. For linear applications such as the grounded base amplifier, the Class A power gain is approximately proportional to $a^2 r_c$; hence the gain of such an amplifier will stay essentially constant within a db or two over the temperature range from -40°C to $+80^\circ\text{C}$. For pulse applications, and of importance to de biasing with point-contact transistors, is the fact that the de collector current (for fixed emitter current and collector voltage) will change at about the

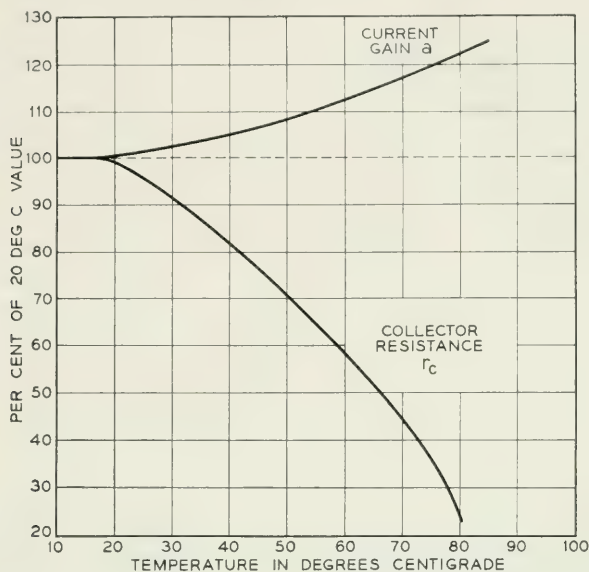


Fig. 22—Collector resistance and a versus temperature for type A transistor

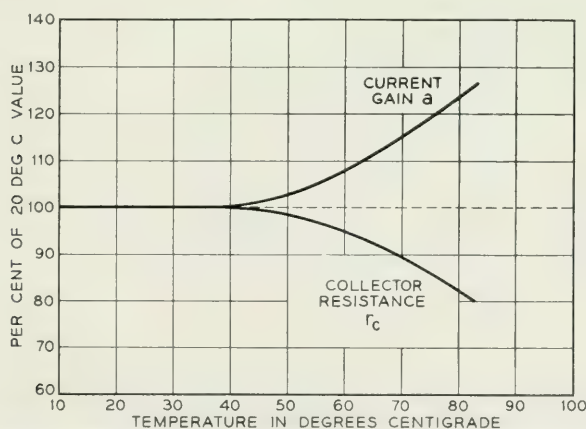


Fig. 23—Collector resistance and a versus temperature for type M1729 transistor.

same rate as does r_c , the small signal collector impedance. Similar improvements have been made in these variations for switching transistors and Fig. 24 is a series of graphs showing how the M1689 bead type switching transistor changes the pulse characteristics defined in Fig. 20 with respect to temperature. For those switching functions examined to date, it is believed that these data mean reliable operation to as high as $+70^\circ\text{C}$ in most applications and perhaps as high as $+80^\circ\text{C}$ in others.

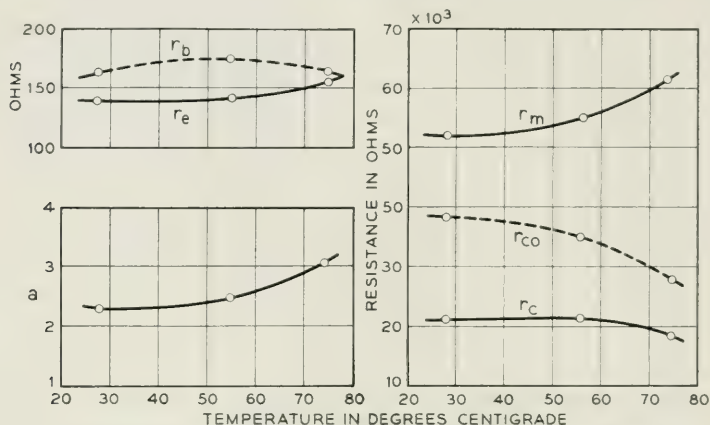


Fig. 24—Temperature behavior of the M1689 transistor.

In junction transistors the laws of temperature variation are not so well established, the device being in a much earlier stage of development. Preliminary data indicate smaller variations in the small signal parameters such as a and r_e . On the other hand, variations in the dc current, particularly I_{co} , are many times greater, of the order of 10 per cent per degree centigrade.* The only saving grace here is the fact that I_{co} is normally very much less than the actual operating value of I_e .

In summary, it may be said that while significant improvements have been made in temperature dependence to the point where many applications appear feasible, it is not to be inferred that the temperature limitation is completely overcome. Much more development work of device, circuit and system nature is required to bring this aspect of reliable operation to a completely satisfying solution.

Shock and Vibration

With regard to mechanical ruggedness, current point-contact transistors have been shock tested up to 20,000 g with no change in their

* I_{co} is the collector current at zero emitter current.

electrical characteristics. Vibration of point-contact and junction transistors over the frequency range from 20 to 5000 cps at accelerations of 100g produces no detectable modulation of any of the transistor electrical characteristics, i.e., such modulation, if it exists, is far below the inherent noise level. At a few spot frequencies in the audio range, vibration tests up to 1000g accelerations similarly failed to produce discernible modulation of the transistor characteristics.

MINIATURIZATION FIGURE OF MERIT	TYPE A SEPTEMBER 1949	JANUARY 1952	NEW DEVELOPMENT TYPE
VOLUME	$\frac{1}{50}$ IN ³	$\frac{1}{2000}$ IN ³	POINT - M1689
		$\frac{1}{500}$ IN ³	JUNCTION - M1752
MINIMUM COLLECTOR VOLTAGE FOR CLASS A OPERATION	30 V	2 V	POINT - M1768, M1734
		0.2 V	JUNCTION - M1752
MINIMUM COLLECTOR POWER FOR CLASS A OPERATION	50 MW	2 MW	POINT - M1768
		10 μ W	JUNCTION - M1752
CLASS A EFFICIENCY	20 %	35 %	POINT - M1768, M1729
		49 %	JUNCTION - M1752

Fig. 25—Miniaturization in space and power drain.

MINIATURIZATION STATUS

Space Requirements

In smallness of size, the transistor is entering new fields previously inaccessible to electron devices. The cartridge structure (see Fig. 25), such as the type A, has a volume of $\frac{1}{50}$ cubic inch, compared to about $\frac{1}{8}$ cubic inch for a sub-miniature tube and about 1 cubic inch for a miniature tube. Under current development, the M1689 bead point-contact transistor has substantially similar electrical characteristics to the M1698* cartridge switching unit but occupies only about $\frac{1}{2000}$ cubic inch. The M1752 junction bead transistor has a volume of approximately $\frac{1}{500}$ cubic inch but this may be reduced to the same order as the point-contact bead if necessary. For further substantial size reductions in equipment, the next move must comprise the passive components. It should be pointed out that the low voltages, low power drain, and correspondingly lower equipment temperatures should make possible further reductions in passive component size.

* The M1698 transistor is a cartridge type point-contact transistor with electrical characteristics designed for switching and pulse applications. This unit is proving useful in the laboratory development of new circuits or in cases where miniature packages are unnecessary.

Power Requirements

The transistor, of course, has the inherent advantage of requiring no heater power; moreover, significant advances have been made in the past two years in reducing the collector voltage and power required for practical operation. Consider the minimum collector voltage for which the small-signal Class A gain is still within 3 to 6 db of its full value. In September, 1949, the type A transistor could give useful gains at collector voltages as low as 30 volts. Today, several point-contact devices (M1768 and M1734) perform well with collector voltages as low as 2 to 6 volts even for relatively high-frequency operation. One junction transistor, the M1752, can deliver useful gains at collector voltages as low as 0.2 to 1.0 volt. Under these same conditions, the minimum collector power for useful gains may be as low as 2–10 mw for point-contact devices and as low as 10 to 100 μ w in the case of the junction transistors.* Class A efficiencies have been raised for the point-contact devices to as high as 30–35 per cent and for junction transistors this may be as high as 49 per cent out of a maximum possible 50 per cent. Class B and C efficiencies are correspondingly close to their theoretical limiting values.

PERFORMANCE STATUS

Exact electrical performance specifications for the transistor depend, of course, upon the intended applications and the type of transistor being developed for such an application. These types are beginning to be specified; and in fact, they are already so numerous that mention of only a few salient features of some of them will be attempted. Bear in mind, as was pointed out before, that no one transistor combines all the virtues any more than does any one tube type. Fig. 26 attempts to compare the progress made in several important performance merit figures by development of several point-contact and junction types during the last two years. Again the reference performance is that of the type A as of September, 1949.

Some switching and transmission applications need transistors having high current gain. By going to a point-junction structure, useful values of alpha as high as 50 are now possible with laboratory models.

For straight transmission applications, the single stage gain of point-contact types (M1768, M1729) has been increased to 20–24 db, whereas for the M1752 junction type the single stage gain may be as high as 45–50 db.

* In some special cases, depending upon the application, practical operation may be obtained for as little as 0.1 to 1.0 microwatt.

PERFORMANCE FIGURE OF MERIT	TYPE A SEPTEMBER 1949	JANUARY 1952	NEW DEVELOPMENT TYPE
α - CURRENT GAIN	5 X	50 X	JUNCTION
SINGLE STAGE CLASS A GAIN	18 DB	22 DB	POINT - M1729, M1768
		45 DB	JUNCTION - M1752
NOISE FIGURE AT 1000 CPS	60 DB	45 DB	POINT - M1768
		10 DB	JUNCTION - M1752
FREQUENCY RESPONSE f_c	5 MC	7-10 MC	POINT - M1729
		20-50 MC	POINT - M1734
CLASS A POWER OUTPUT	0.5 WATT	2 WATTS	JUNCTION
SWITCHING CHARACTERISTICS	NONE	GOOD	POINT - M1698, M1689 M1734
FEEDBACK RESISTANCE r_b	250 OHMS	70 OHMS	POINT - M1729
LIGHT DARK PHOTOCURRENT RATIO	2:1	20:1	JUNCTION - M1740

Fig. 26—Performance progress.

For high-sensitivity low-noise applications, the point-contact devices have been improved to have noise figures of only about 40–45 db, whereas the M1752 *n-p-n* transistor has been shown to have noise figures in the 10–20 db range. All such noise figures are specified at 1000 cps and it should be remembered that they vary inversely with frequency at the rate of about 11 db per decade change in frequency.

For video, I.F., and high-speed switching applications, measurable improvement has been attained in the frequency response. For video amplifiers up to about 7 mc, the M1729 point-contact transistor is capable of about 18-20 db gain per stage. For high-frequency oscillators and microsecond pulse switching, the M1734 point-contact transistor is under development. Preliminary models of 24 mc I.F. amplifiers using the M1734 have been constructed in the laboratory, these amplifiers having a gain of some 18–24 db per stage and a band-width of several megacycles. However, more work needs to be done on the M1734 to reduce its feedback resistance. For pulse-handling functions, such M1734 units work very nicely as pulse generators and amplifiers of $\frac{1}{2}$ micro-second pulses, requiring only 6–8 volts of collector voltage and 12–20 mw of collector power per stage. The amplified pulses can have ampli-

tudes as large as 4–5 volts out of a total collector voltage of 6 volts and rise times as little as 0.01–0.02 microsecond.

By increasing the thermal dissipation limits of junction transistors, the Class A power output has been raised to 2 watts in laboratory models. This, however, does not represent an intrinsic upper limit but rather a design objective for a particular application.

Characteristics suitable for switching are now available in the M1698, M1689 and M1734 point-contact types, as previously described, but this is a continually evolving process and more work certainly remains to be done. At present it is possible to operate telephone relays requiring as much as 50 to 100 ma with M1689 and M1698 point-contact transistors.

New junction-type phototransistors⁵ represent a marked advance over the earlier point-contact type.⁶ While their quantum efficiencies are not as high as those of the point-contact types, nevertheless the light/dark current ratios are greatly improved and the collector impedance has been raised 10–100 times thus making possible much greater output voltages for the same light flux.

SOME SELECTED APPLICATIONS

Data Transmission Packages

To determine the feasibility of applying transistors in the form of miniature packaged circuit functions, several of the major system functions of a pulse code data transmission system have been studied. This investigation has been undertaken under the auspices of a joint services engineering contract administered by the Signal Corps.

It was desired that these studies should lead to the feasibility development of unitized functional packages combining features of miniaturization, reliability and lower power drain. Accordingly, it was necessary to carry on in an integrated fashion activities in the fields of system, circuit and device development to achieve these ends. In particular, circuit and system means have been developed to perform with transistors the functions of encoding, translation, counting, registering and serial addition. The M1728 junction diode, M1740 junction photocell and M1689 bead switching transistor are direct outgrowths of this program and are the devices used in the circuit packages.

At this point, the major system functions shown in Fig. 27 have been achieved with interchangeable transistors. These major system functions are in turn built up of some seven types of smaller functional packages listed in Fig. 28. The end result of this exploratory development can be

said to have demonstrated the feasibility of such a data transmission system in the sense that a workable (though not yet optimal) system can be synthesized from reproducible transistor-circuit packages which have been produced at reasonable yields and with reasonable (though not yet complete) service reliability. Further development work would be needed in all phases to make such a system of packages suitable for field use. It is estimated that the present laboratory model requires about one-tenth the space and power required to do the same job with present tube art. Fig. 29 is a photograph of a transistor bit-register package and Fig. 30 is another photograph of such packages showing both sides of the various types employed.* Actual final packages would

1. 4 DIGIT REVERSIBLE BINARY COUNTER
2. 6 DIGIT ANGULAR POSITION ENCODER
3. 6 DIGIT GRAY-BINARY TRANSLATOR
4. 5 DIGIT SHIFT REGISTER
5. 2 WORD SERIAL ADDER

Fig. 27—System functions tested.

DEVELOPMENT PACKAGE TYPE	PACKAGE FUNCTION	DEVELOPMENT TRANSISTOR, DIODE TYPES USED
M 1731-1	REGENERATIVE GATE	M 1689 M 1727
M 1732-1 M 1736 M 1790	BIT REGISTER	M 1689 M 1727 M 1734
M 1733-1 M 1792	PULSE AMPLIFIER	M 1689
M 1735-1 M 1747-1 M 1748-1 M 1751-1 M 1751-2 M 1751-3	DIODE GATE	M 1727 400A
M 1745-1 M 1791	BINARY COUNTER	M 1689 400A
M 1749-1	PHOTOCELL READOUT	M 1740
M 1746-1	DELAY AMPLIFIER	M 1689

Fig. 28—Development transistor—circuit packages.

* The Auto-Assembly Process used in the construction of these packages is a Signal Corps Development.



Fig. 29—Bit register package.

probably not use such clear plastics and Fig. 31 shows some packages in which the plastic has been loaded with silica to increase its strength and thermal conductivity. The assembly in Fig. 31 consists of a six-digit position encoder at the left, followed by six regenerative pulse amplifiers which in turn feed a six-digit combined translator-shift register.

*N-P-N Transistor Audio Amplifier and Oscillator**

To the right in Fig. 32 is shown a transformer-coupled audio amplifier employing two M1752 junction transistors. This amplifier has a pass band from 100–20,000 cps and a power gain of approximately 90 db. Its gain is relatively independent of collector voltage from 1–20 volts,

* The material of this section represents a summary of some work by Wallace and Pietenpol described more completely in Ref. 4.

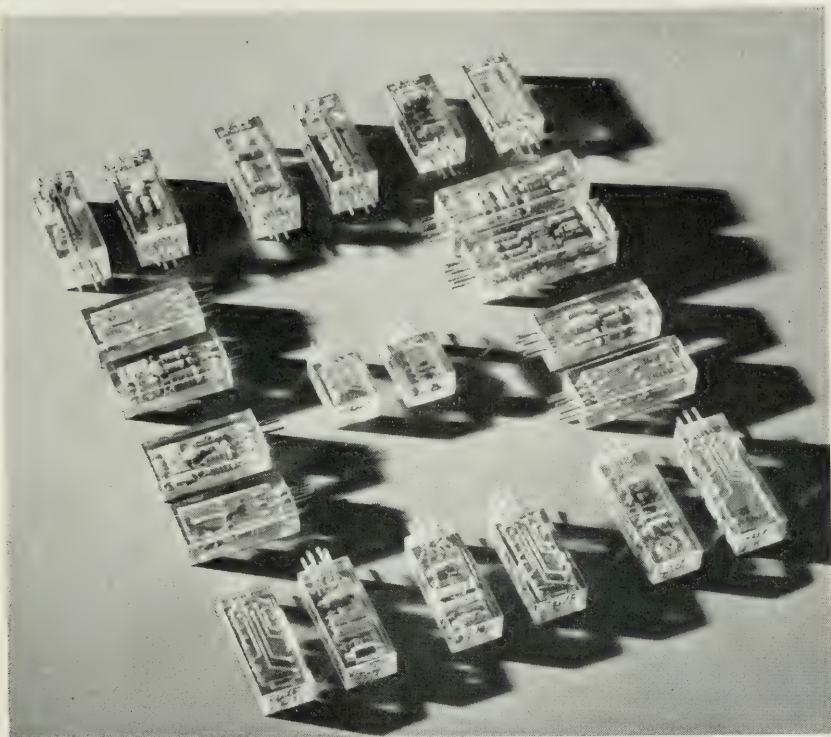


Fig. 30—Package construction illustrated.

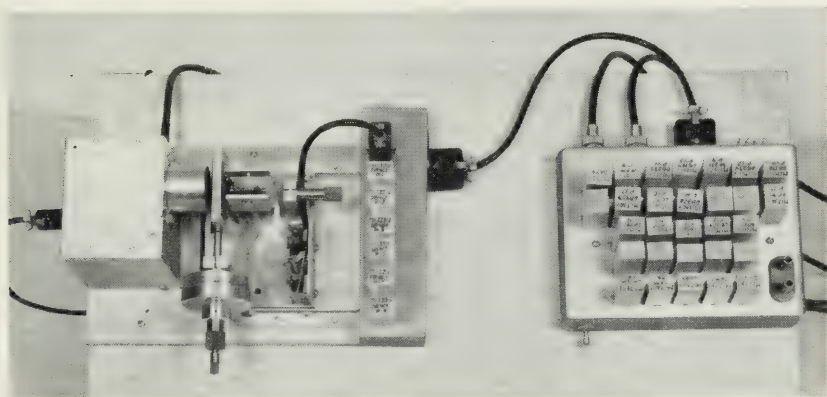


Fig. 31—Laboratory model of encoder-transistor-register using transistor packages.

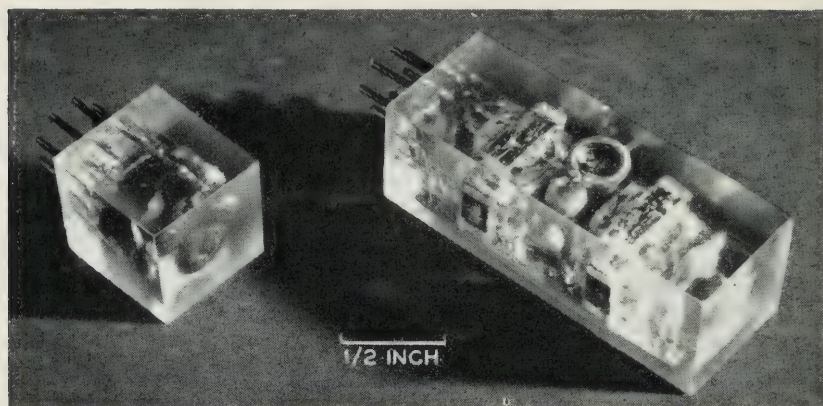


Fig. 32—Packaged oscillator and amplifier using junction transistors.

only the available undistorted power output increasing as the voltage is increased. At a collector voltage of 1.5 volts it draws a collector current of approximately 0.5 ma per unit for a total power drain of 1.5 milliwatts. Under these conditions it will deliver Class A power output of about 0.7 milliwatt. The noise figure of such an amplifier has been measured to be in the range from 10–15 db at 1000 cps depending upon the operating biases.

To the left of Fig. 32 is shown a small transistor audio oscillator having a single M1752 transistor, a transformer and one condenser. To see just how little power was the minimum necessary to produce stable oscillations such an oscillator was tried at increasingly lower collector supply voltages. It was found that stable oscillations could be maintained down to collector supply voltages as low as 55 millivolts and collector current as low as 1.5 microamperes for a total drain of 0.09 microwatt.

SUMMARY

With respect to reproducibility and interchangeability, transistors now under development appear to be the equal of commercial vacuum tubes.

With regard to reliability, transistors apparently have longer life and greater mechanical ruggedness to withstand shock and vibration than most vacuum tubes. With regard to temperature effects, transistors are inferior to tubes and present upper limits of operation are 70–80°C for most applications. This restriction is often reduced in importance by the lower power consumption which results in low equipment self-heating. This, however, is the outstanding reliability defect of transistors.

With regard to miniaturization, the comparison figures are so great as to speak for themselves. Operation with a few milliwatts is always feasible and in some cases operation at a few microwatts is also possible.

With regard to performance range, it is believed that the above results imply the following tentative conclusions:

In pulse systems (up to 1–2 mc repetition rates) transistors should be considered seriously in comparison to tubes, since they provide essentially equal functional performance and have marked superiority in miniature space and power. Bear in mind that in some reliability figures they are superior whereas in the matter of temperature dependence they are inferior to tubes.

In CW transmission at low frequencies (< 1 mc) essentially the same conclusions are indicated, primarily because of junction transistors. In the range from 1–100 mc, tubes are currently superior in every functional performance figure (except perhaps noise and bandwidth) so that for transistors to be considered for such applications, much greater premium must be placed on miniaturization and reliability than for the first two applications areas.

Thus, it might be assumed that, even though there are many outstanding development problems of a circuit and device nature to be solved, it is appropriate for circuit engineers to explore seriously the application possibilities of transistors—not only in the hope of building better systems, but also to influence transistor development towards those most important systems for which their intrinsic potentialities best fit them. It should not be inferred that all important limitations have been eliminated—nor, on the other hand, that the full range of performance possibilities have been explored.

If one remembers the history of engineering research and development in older related fields, it seems apparent that a relatively short time has elapsed since the invention of the first point-contact transistor. Already, new properties and new types of devices are under study and some have been achieved in the laboratory. It therefore is possible, and certainly stimulating, to infer that more than a single new component is involved; that much more lies ahead than in the past; that, indeed we may be entering a new field of technology, i.e., “transistor electronics”.

ACKNOWLEDGMENTS

It was stated earlier that these advances in the development of transistors have resulted from improved understanding, materials and processes. These improvements have been made through the efforts of a large

number of workers in physical research, chemical and metallurgical research and transistor development. In reality, these colleagues are the authors of this paper; and it is to them the writer owes full and appreciative credit for the material that has made possible this report of progress in transistor electronics.

REFERENCES

1. R. M. Ryder, R. J. Kircher, "Some Circuit Aspects of the Transistor", *Bell System Tech. J.*, **28**, p. 367, 1949.
2. R. L. Wallace, G. Raisbeck, "Duality as a Guide in Transistor Circuit Design", *Bell System Tech. J.*, **30**, p. 381, 1951.
3. W. Shockley, M. Sparks, G. K. Teal, "*p-n* Transistors", *Phys. Rev.*, **83**, p. 151, 1951.
4. R. L. Wallace, W. J. Pietenpol, "Some Circuit Properties and Applications of *n-p-n* Transistors", *Bell System Tech. J.*, **30**, p. 530, 1951.
5. W. J. Pietenpol, "*p-n* Junction Rectifier and Photocell", *Phys. Rev.*, **82**, No. 1, pp. 122-121, Apr. 1, 1951.
6. J. N. Shine, "The Phototransistor", *Bell Laboratories Record*, **28**, No. 8, pp. 337-342, 1950.

An Experimental Electronically Controlled Automatic Switching System

By W. A. MALTHANER AND H. EARLE VAUGHAN

(Manuscript received February 15, 1952)

An automatic telephone switching system, built as a laboratory experiment, is described in which electronic techniques, high speed relays and a subscriber telephone with a pre-set dialing mechanism were employed. One-at-a-time operation within the office was made possible by these fast tools; that is, only a single control circuit was provided for each function. This experimental system, although not commercially economical, showed that an advantageous reduction in the number of control and connector circuits is made possible by this method of operation.

INTRODUCTION

This paper describes a laboratory experiment in automatic telephone switching systems. The investigation was conducted at the research level to gather valuable information and circuit techniques from a laboratory trial and not to evolve a system economically competitive with existing systems since the area of investigation is always broader and the results more general in character when the work is unfettered by economic restraints. Indeed, the results are not economically competitive.

Purposes of the investigation were to determine what advantages may be derived from faster operation, largely through the use of electronic techniques, and to introduce and test some previously unexplored philosophies in switching and signaling. Some of the basic tools employed were dry-reed relays, mercury relays, multi-element cold cathode gas tubes, cold cathode gas diodes, and thermionic electron tubes. An experimental subscriber's telephone set, incorporating a preset dial mechanism with circuits for generating dialing signals of a new form, together with suitable signal receivers for the central office was designed as well as a novel type of switching network with its control circuits. A basic aim of the experiment was one-at-a-time operation within the central office.

BACKGROUND AND OBJECTIVES

In many recent designs of dial telephone central offices, especially those in use in large urban areas, the subscriber's dial does not control directly the setting of switches leading toward the desired destination as was the case in early dial systems. Instead the information is received first by a register circuit which is selected from a group of such register circuits and is connected to the calling subscriber's line on the origination of a call. The register cooperates with other complex circuits to ascertain the location of idle trunks to the called subscriber's office and possible routes through the switching network to these trunks, and to control the selection and use of one such path to this called office. In the called office another register circuit, frequently of a type different from that into which the subscriber originally dialed, is selected from a group of such circuits and the directory number of the called subscriber is transmitted to it from the register-sender circuit in the calling office. In the terminating office the procedure of locating and testing the called line and switching paths to it, and of establishing a connection over one of these paths is accomplished through the use of additional control circuits. These various circuits which are used in setting up a conversational path are called common control circuits.

Each type of common control circuit is provided in sufficient number to handle the expected traffic. The number required is, of course, related to speed of operation since the shorter the holding time of a circuit, i.e., the length of time a circuit takes to complete its functions for one call, the more calls such a circuit can complete in a given time. The holding time of a control circuit is, in turn, dependent upon the operating speed of the equipment controlled. Furthermore, control circuits of the same type, if more than one of a given type is required, will have added to their normal functioning time during busy traffic periods a delay time interval since they must not interfere with each other's actions in the controlled equipment. Common control circuits, such as dial pulse registers, which receive information directly from subscribers must be engineered on the basis of an average holding time which allows for the variable reaction times, hesitations, partial usages and other personal idiosyncrasies of subscribers. Present designs of automatic central offices require a number of each type of control circuit and auxiliary circuits for selecting and connecting the control circuits as required in the operation of the system. These control circuits and connectors embrace a considerable fraction of the space and cost of such an office.

Dr. T. C. Fry, at the time he was Director of Switching Research

at the Bell Telephone Laboratories, suggested that a program be started to explore the possibilities of a new system which would require only a single control circuit of each type. This would require that each group of functions assigned to a common control circuit be performed on a one-call-at-a-time basis. It might be accomplished in a fresh approach to system design employing recent developments in high speed components. High speed in the common control units alone would not be sufficient. It would also be necessary to have fast switches since the operating time of a switching network is part of the holding time of the control circuit which operates the network. Similarly, since the signaling time is part of the holding time of the control circuit which receives and registers the signals, some form of high speed signaling would also be required. Further, the subscribers should have no direct control of the holding time of any common control unit. It was hoped that a great reduction in the number of common control circuits and connectors would result in a reduction in the size and cost of a central office even if the individual control circuits were somewhat more expensive. Furthermore, a speed permitting one-at-a-time operation would result automatically in faster service for the subscriber.

Consideration of the various factors of one-at-a-time operation was undertaken by the members of the Switching Research Department and possible system components evolved. Primary elements of inherently high speed, such as cold cathode gas tubes, thermionic electron tubes, dry-reed relays and mercury relays, were immediately adopted for the system. A network of high-speed switches with its high-speed control circuits was designed. A pre-set dialing device in the subscriber telephone set with transmission of high-speed dialing signals to the central office under control of common equipment in the office was selected as a means of eliminating the direct influence of subscribers on control circuit holding time. A code of high-speed signals, suitable for transmission over all existing types of local telephone facilities, with means for the pre-selection and controlled generation of telephone numbers was designed into the subset. Such a subset is necessarily complex since it becomes a form of manually operated register with all digits of a number stored before transmission to the central office. Circuits to control the generation of subset signals from the central office and receiver circuits to decode and register the signals were constructed.

These parts were then combined in the design of the Electronically Controlled Automatic Switching System, ECASS. A skeletonized laboratory version was built and tested to investigate the feasibility of combining the circuit elements and techniques, and to prove the operability

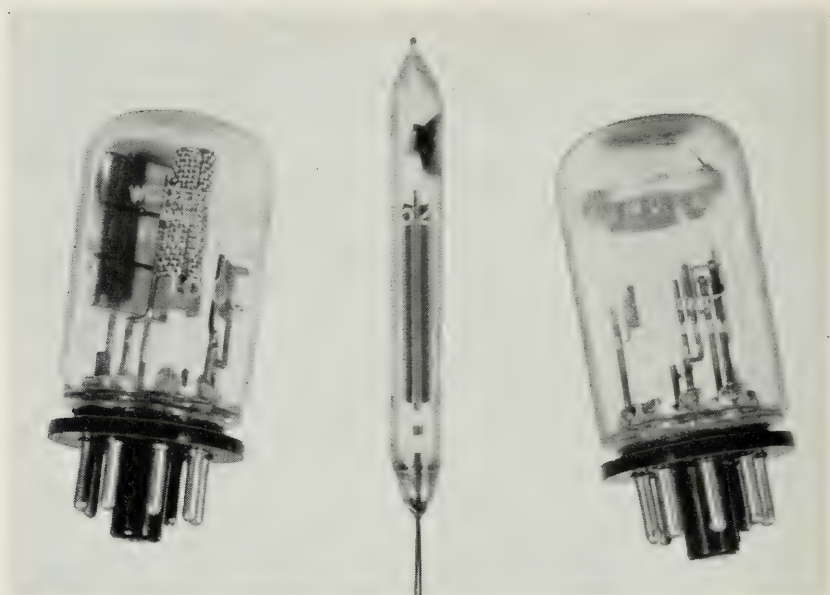


Fig. 1—Cold-cathode gas tubes—pentode, diode and octode.

of such a system. System operation is described in this paper after a more detailed discussion of the components mentioned above.

COMPONENTS

Cold cathode tubes, usually diode or triode types, have found widespread application in the past but the gas tubes used in the ECASS system were developed to have special characteristics for switching use. The three types of cold cathode gas tubes used were: a diode, a screen grid pentode and a multi-purpose octode. Fig. 1 shows a photograph of each type and Fig. 2 gives a schematic drawing of the internal elements. These tubes were developed by W. A. Depp and R. L. Vance. The diode

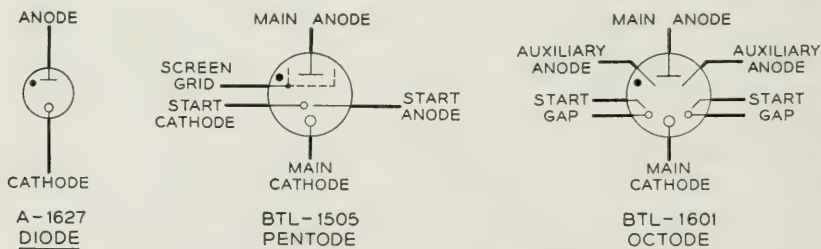


Fig. 2—Schematics of cold-cathode gas tubes.

is used at many points throughout the switching network, the screen-grid pentode in the path selection processes in the switching network, and the octode for miscellaneous purposes in the line, trunk, number group and other circuits.

The dry-reed switch, which is used as the contact element in many fast relays as well as in the metallic talking path through the office, is shown in Fig. 3. This switch consists of two permalloy rods sealed in opposite ends of a small glass tube which is filled with an inert gas. The overlapping ends of the rods normally have a gap between them and the application of a magnetic field coaxial with the reeds will cause them to pull together and close a metallic path from one rod or reed to the other through rhodium plating at the contacting ends. The dry-reed switch has an extremely small operate and release time, and because of the gas sealed and permanently adjusted construction provides a highly reliable dirt-free contact for low current applications. The dry-reed switch and relays employing it were developed by W. B. Ellwood. Mercury contact relays, also of a sealed and permanently adjusted construction, are used where fast operation at heavier currents is required. A sectional drawing

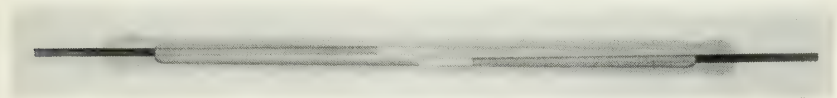


Fig. 3—Glass-sealed dry-reed switch.

of a mercury contact relay is shown in Fig. 4. These relays were developed by J. T. L. Brown and C. E. Pollard. Dry-reed relays and mercury relays are described in *Electrical Engineering*, Vol. 66, pp. 1104–1109, November, 1947, and in *Bell System Monograph*, 1516.

THE PRE-SET SUBSCRIBER'S TELEPHONE

In order to eliminate direct control of any common equipment by the subscriber and thereby to reduce the holding time of the dialed information receiving circuits and the associated subscriber-connecting circuits, the experimental pre-set dial telephone set shown in Fig. 5 was designed for this system by K. S. Dunlap, H. E. Hill and D. B. Parkinson. Eight selector finger wheels are grouped on a common shaft with only their edges visible across the front of the telephone housing. Each finger wheel is provided with ten indentations along its exposed periphery. Each indentation is designated by an engraved number or group of letters conforming to the telephone directory numbering system and each indentation is of suitable configuration to permit a subscriber's

finger to engage and move the wheel in either direction to one of ten detented positions. All of the wheels may be returned to their normal "zero" position simultaneously by depressing the release button on the front right corner of the housing. To place a call the subscriber positions each of the wheels so that the desired number may be read across the wheels on the line of indentations immediately above the lower edge of the enclosing frame. The first three wheels are set to the code of the called office and the next five to the called line directory number with the last of these being used for the party letter, if required. A number is preset in this manner before the handset is removed from its cradle across the back of the housing. With this method of operation the number may be rapidly and completely transmitted to the central office when its receiving circuit has been connected to the line.

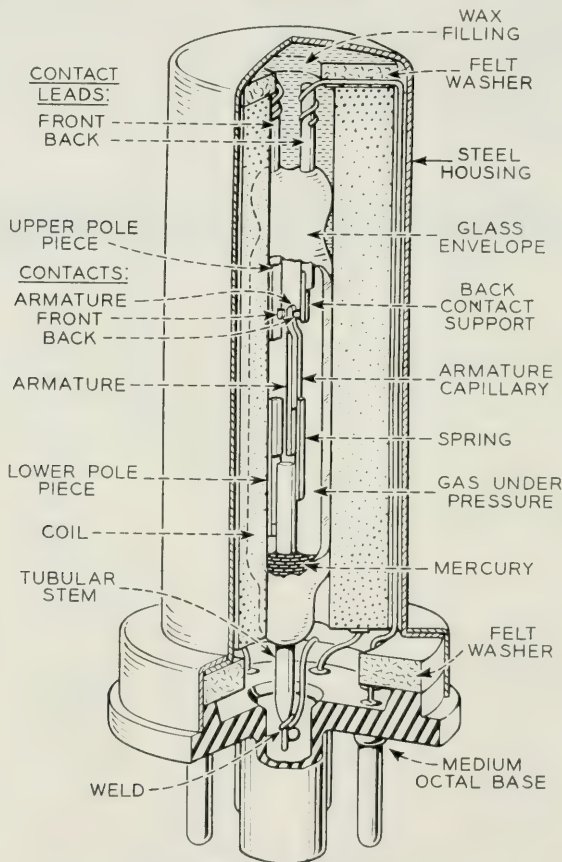


Fig. 4—Mercury contact relay.



Fig. 5—Pre-set pulse-position-dialing telephone set.

As shown in Fig. 6, which is a schematic of the mechanism and circuit of this telephone set, the handset when resting in its supporting cradle depresses the switchhook pins and causes two bell cranks to operate two sets of switchhook contact assemblies. One of these contact assemblies is controlled solely by the position of the handset while the other contacts are controlled jointly by the handset and by a magnetic locking device. This magnetic locking device consists of a permanent magnet yoke which holds the contacts in the position shown after the removal of the handset from its cradle until direct current of the correct polarity is allowed to flow in the windings of a latch magnet.

These two sets of switchhook contacts jointly control the connection of any of three subdivisions of the apparatus in the telephone set to the line to the central office. If the handset is removed from its cradle to originate a call, the free set of switchhook contacts releases to complete a circuit through the latched set of contacts to the signaling equipment of the station. In this signaling condition the voice transmission equipment remains disconnected from the circuit; thus, interference and transmission losses caused by voice transmission equipment are avoided during signaling. Upon completion of signaling direct current is provided from the central office to trip the latched switchhook contacts.

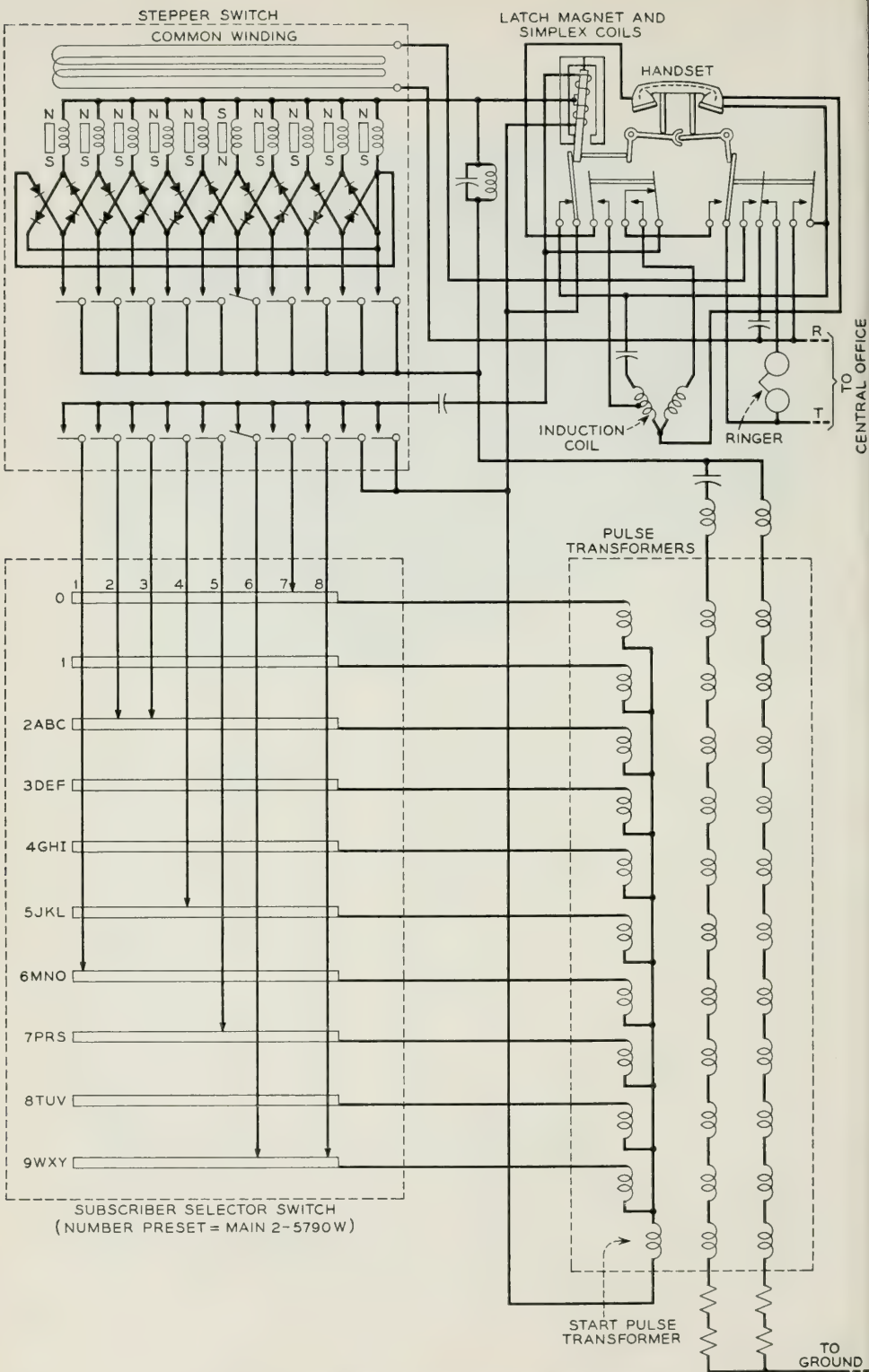


Fig. 6—Pulse-position-dialing subset schematic.

With both sets of switchhook contacts now released the usual transmitter, receiver and induction coil arrangement for transmission of voice currents is connected to the telephone line and all of the station signaling equipment, including the tripping windings of the latch magnet, is disconnected from the circuit. Interference and transmission losses caused by signaling equipment are thus avoided during conversation. When the handset is resting on its cradle between calls with both sets of switchhook contacts operated, the usual ringer and ringer condenser are connected across the line for responding to incoming calls. Upon removal of the handset in answer to such an incoming call, direct current is provided from the central office to trip the latched switchhook contacts and thereby the set is placed immediately in the talking condition.

PULSE POSITION DIALING SIGNALS

Before describing further the operation of this telephone set, it will be necessary to explain briefly the dialing signals generated by it and used in the system.

From the subscriber's telephone set eight digits are transmitted for a complete local area directory number and the transmission is repeated as many times as necessary for the functioning of the central office equipment. In order to indicate the starting point of the transmission of a complete called number, a time interval of two digits duration during which no signals are transmitted is provided at the beginning of each transmission. Each digit interval is 0.01 seconds; therefore, a time interval of 0.1 second is required for the no-signal or blank period and the eight digit number.

These signals, as shown in the wave form-time diagrams of Fig. 7, consist of two pulses per digit: a start pulse of 1 millisecond duration and a stop pulse of 1 millisecond duration, each pulse approximately a single cycle of a 1,000-cycle per second sine wave. The time interval between a start pulse and its following stop pulse is the measure of the associated digit value. The start pulses are generated at intervals of 0.01 seconds, or 10 milliseconds, and one stop pulse is generated some time during the 3.2 to 6.8 millisecond interval after each start pulse. In order to provide sufficient margins to permit reliable signaling over a wide variety of transmission facilities 3.2 milliseconds are allowed for the decay of each pulse and the pulses themselves occupy a section of the voice-frequency spectrum transmitted by practically all communication facilities. The possible starting times of stop pulses representing

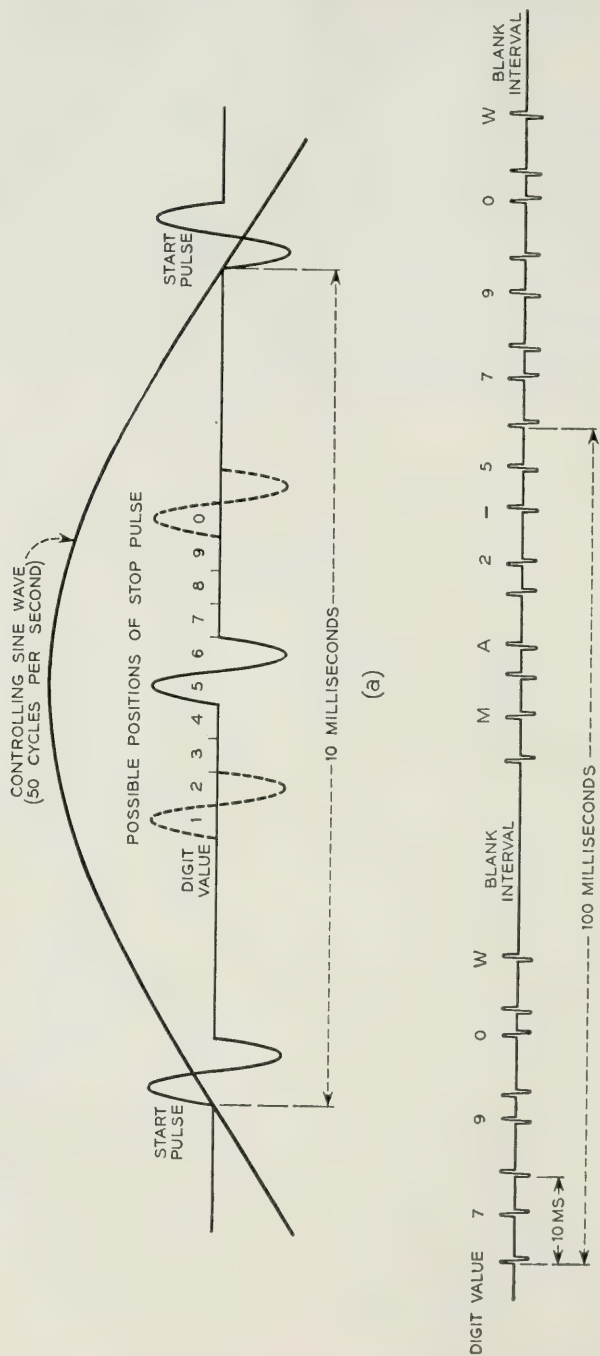


Fig. 7—Pulse-position-dialing signals.

digits of successive magnitudes differ by 0.4 milliseconds. Thus, digit 1 is represented by a start pulse followed by a stop pulse 3.2 milliseconds later; digit 2 is represented by a start pulse followed by a stop pulse 3.6 milliseconds later; and so on. It will be observed that the stop pulse for the digit 0 is 6.8 milliseconds after its start pulse and 3.2 milliseconds before the next succeeding start pulse. Thus, there is provided an increment of time of 3.2 milliseconds for the decay of the start pulse, increments of 0.4 milliseconds each for the generation of a pulse at any one of the ten times necessary to represent the various digits, and a last increment of 3.2 milliseconds to permit a stop pulse to decay should it occur at the end of the ninth increment of time.

Referring again to Figure 6, the signaling pulses are generated by the eleven pulse transformers shown. These saturation-type transformers are assigned, one for each of the numerals 0 to 9 and one for the start pulse. The excitation for the signaling apparatus is a constant amplitude 50-cycle current of sinusoidal wave form transmitted from the central office on a simplex circuit consisting of the two line wires to the set with ground return.* The currents from the line wires pass into the signaling apparatus through the windings of the latch magnet. These latch magnet windings thus serve also as a simplex coil and since the excitation magneto-motive-forces in the two windings are mutually opposing there is no reaction on the latch itself.

From the simplex coil the excitation current flows through a stepper switch and its shunting phase shifter to a phase splitting network in which the current is converted to a two phase source with its two currents 90 degrees out of phase. Each of the pulse generating transformers has a single secondary and two primary windings. The primary windings of the transformers are serially interconnected and connected with the two phases of the excitation current so that one phase is applied to one primary winding of each transformer and so that the other phase is applied to the other primary winding of each transformer. The secondary windings are connected across the line through the pre-set selector, contacts of the stepping switch and a series capacitor. The secondary winding of the pulse transformer for the start pulse is in a lead common to all the stop pulse secondaries.

The magnetic core of each pulse transformer is designed to be saturated except for very small values of ampere-turns, and a voltage pulse

* The time interval spacings of signal pulses given in this section and in the following section on the signal receiver are based on a 50-cycle control current. The system operated satisfactorily on 50 cycles. However, in most of the laboratory tests a control current of 45 cycles per second was used since a stable source of this frequency is readily derived from commercial 60-cycle power sources.

is generated in the secondary winding of each transformer when the flux is changed from saturation at one polarity to saturation at the other polarity. The flux generated in the core of each transformer depends upon the number of turns in the two primary windings and upon the current flowing in each winding. In order to assure that all pulses be substantially alike as to wave form and amplitude it is necessary that the total maximum ampere-turns on each core be equal. In order to cause each transformer to generate a pulse at a suitable time during each half-cycle of the excitation current the total ampere-turns driving flux through the transformer cores must be controlled so that the flux in each transformer is zero at the time assigned to the pulse which that transformer serves to generate. These conditions determine the number of turns and the polarity of each winding when the angular position of the desired pulse is fixed in relation to each half-cycle of the basic excitation current.

Since the magnetic flux in each transformer is reduced to zero two times during each cycle of excitation current, it follows that a combination of two pulses representing a digit must occur during each half-cycle of the excitation current and that each combination of two pulses representing a digit is of opposite polarity to the preceeding two pulses. The capacitor through which the pulse generating transformer secondary windings are connected to the line is so proportioned to the impedances of these windings and to the impedance of the line that each half-cycle pulse as generated by a transformer is applied to the line as a single complete cycle of alternating current of about 1 millisecond duration.

A selector switch, which is the internal mechanism connected with the finger wheels pre-set by the subscriber, serves to interconnect the transformer pulse windings with the line through the stepper switch. Thus, pulses representing any of the digits 0 to 9 may be impressed across the telephone line as any desired part of a complete telephone number in accordance with the setting of the selector switch.

The stepper switch employs ten relays of the glass-sealed dry-reed type and each of the relays has an individual coil surrounding two normally open reed contacts. The reeds are polarized by a permanent magnet of sufficient strength to hold the reed contacts closed but not strong enough to close them until assisted by current of the correct polarity through the winding. A reverse current through the winding is required to release the contacts. In addition a common winding is provided which surrounds all of the reeds in such a manner that when a current of sufficient magnitude is passed through the winding the reeds of a predeter-

mined delay will be closed and the reeds of all the other relays will be opened. This action is produced by reversing the individual winding and bias magnet of the single relay which is to be operated by the current through the common stepper winding. The preliminary setting of the stepper to insure correct operation is provided on each origination of a call by the discharge current from the ringer capacitor through the common winding of the stepper. The ringer capacitor is charged from the central office between calls.

One reed in each of the relays is employed to connect successive brushes of the digit selector switch with the line while the other reed in each relay in conjunction with two diode rectifiers per relay winding is employed to control the operation of the stepper. The stepping operation may be explained by reference to Fig. 6 as follows: The stepper is shown with the reeds for the sixth step closed. When the 50-cycle excitation current makes the terminal common to the individual stepper coils positive with respect to the terminal common to the stepping control contacts, current flows through the upper reed contact of the sixth step, a diode rectifier and the winding of the seventh step relay causing its reeds to close. With the seventh set of reeds closed current flows through a diode rectifier and the winding of the sixth step relay causing its reeds to open. The stepper will remain in this position until the reversal of excitation current a half-cycle later at which time a circuit through an oppositely poled diode rectifier will cause the operation of the relay for the eighth step followed by the release of the relay for the seventh step. The phase of excitation current through the stepper is so adjusted by the shunt phase shifting network that the stepper relays operate and release during the 3.2-millisecond guard interval preceding a start pulse. This prevents mutilation of the signal pulses. The stepping circuit is made reentrant so that the pre-set number will be transmitted repeatedly so long as excitation current is provided.

With the chosen 50-cycle excitation the complete transmission of eight digits and a two-digit silent interval takes only 0.1 second. This results in a short holding time for the central office receiving circuit and the repetitive signaling feature permits repeated trials in case of signal mutilation as well as direct dialing from the subscriber's telephone set to distant offices rather than some form of relayed signaling from registers in the subscriber's own office.

SIGNAL RECEIVER

A simplified block diagram of an experimental receiver for the pulse-position signals used in this system is shown in Fig. 8. The receivers

were designed by N. D. Newby and the authors of this paper. The signals after passing through a bandpass filter are amplified to a standard level by a circuit incorporating backward acting automatic volume control. The arrival of each signal pulse is detected by a threshold device. Since the minimum time interval between the generation of a pulse and the next succeeding pulse is 3.2 milliseconds, the threshold device is arranged to disable itself upon the detection of a pulse for about 3 milliseconds. This prevents false operations of the detector either by tail transients resulting from distortion of a pulse in the transmission medium or by noise occurring in this interval.

When the silent or blank interval which exists between the complete transmission of a number and its next repetition is recognized by the

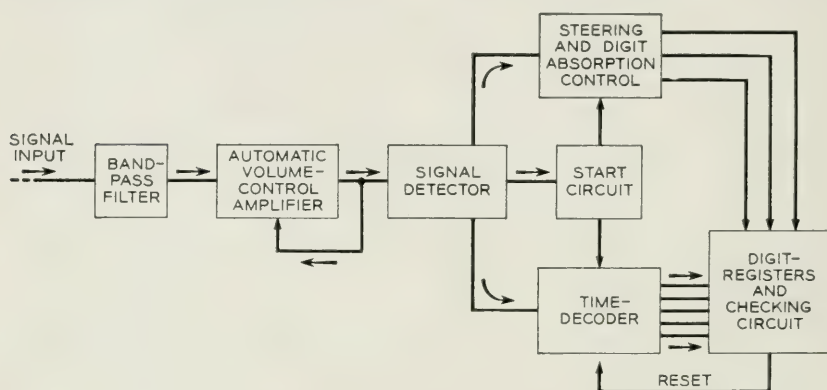


Fig. 8—Pulse-position-dialing receiver.

start circuit attached to the detector, the time-decoder circuit is enabled as well as the steering and digit absorbing circuit. The time decoder subsequently measures the length of time between each detected start pulse and the following detected stop pulse, and energizes the corresponding digit value leads into the registers. The steering circuit enables a separate set of register elements for the storage of each decoded digit which is to be used by its associated circuits and withholds such enablement through its digit absorbing features for digits which are not of immediate interest. The steering circuit also enables a check circuit associated with the registers.

Several features of the signaling code permit a check to be made that the received signals are in accordance with the code. The number transmission cycle has been already described but a brief restatement is made here to emphasize the checkable features: The first pulse following

the blank interval is a start pulse and eight start pulses at uniform 0.01-second time-interval spacing occur between blank intervals. One and only one stop pulse occurs between start pulses. The total number of signal pulses between blank intervals is sixteen. The check circuit utilizes one or more of these properties to insure that no signal pulses have been lost during transmission and that no extraneous pulses have been detected. If the actions of the check circuit indicate that an error in transmission has occurred, the receiver circuits are completely reset for another trial.

THE SWITCHING NETWORK

To meet the objective of a single common control circuit for the operation of the switching network, which provides the selectable paths between any subscriber and any trunk, it was necessary to have the switches in the network considerably faster than any of present commercial design. The laboratory model of the switching network and its associated path selecting equipment employing cold cathode gas tubes and dry-reed relays was developed by E. Bruce and S. T. Brewer. In addition to high operating speed this switching arrangement has certain other desirable properties: The idle path testing and selection functions are incorporated in the internal controls of the network. Busy sections of the network are automatically isolated from the sections tested for subsequent calls. Selection of a trunk within a trunk group, as well as path selection through the network, may be accomplished by the internal controls of the network if the trunks of a group are assigned one trunk per frame. Selection of an idle trunk and an idle switch path in combination reduces blocking. These internal selection controls eliminate many of the connector contacts that would otherwise be required between the switches and external common control circuits.

The switching network consists of line frames and trunk frames with each frame divided into primary and secondary switches. Each primary line and trunk switch has a number of vertical input columns across the switch to which are connected line or trunk circuits respectively and a number of horizontal output rows across the switch. At the intersection of each row and column of a switch is a relay consisting of an operating coil and three dry-reed make contacts. By analogy to the crossbar system which employs a somewhat similar rectangular array of rows and columns per switch and a similar primary-secondary path distribution, a switch intersection is called a crosspoint and a switch relay is called a crosspoint relay. In the crosspoint relay two of the contacts are used

to connect the talking conductors associated with the particular column to the talking conductors associated with the particular row. A cold-cathode gas diode is also associated with each crosspoint relay, and this diode in series with the winding of the relay is connected between the control lead of the particular column and the control lead of the particular row. The third contact of the crosspoint relay is used to short-circuit the associated gas diode. A typical crosspoint is shown schematically in Fig. 9. The use of these crosspoint gas diodes in the control leads facilitates the identification and selection of idle paths through the switching network and the short-circuiting of the diodes at operated crosspoints facilitates the holding of an established connection through the network at a lower power level than required for initial operation and the maintenance of a busy indication along an established connection during the path selection processes of subsequent calls. Dry-reed contact relays, rather than a more conventional type, are used in the crosspoints to provide the operating speed required for single control circuit operation.

Each secondary switch is a similar rectangular array except that the horizontal rows are used as input terminals and the vertical columns as switch outlets. Within a frame the horizontal outputs of the primary switches are interconnected with the horizontal inputs of the secondary

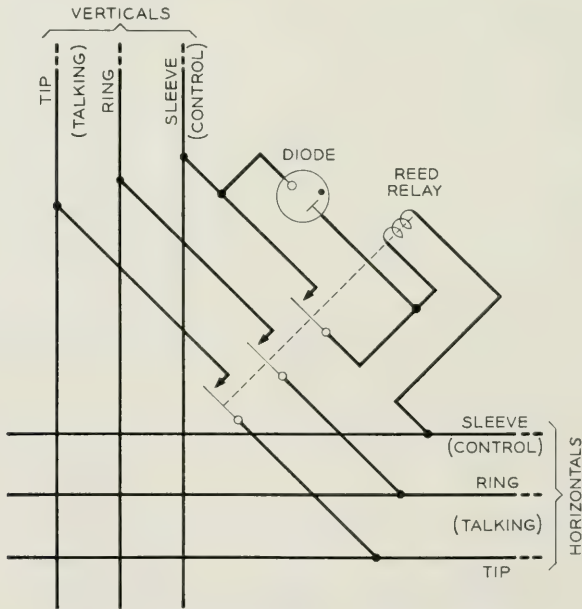


Fig. 9—Reed-diode switch—crosspoint connection.

switches so as to provide one path from each primary switch to each secondary switch.

Connections are made between the secondary line frame switches and the secondary trunk frame switches to provide talking paths between each line frame and each trunk frame. A direct metallic connection is made for the two talking conductors of each path but the control lead from each secondary line switch outlet is connected to an individual control circuit, called a junctor, and the control lead from a secondary trunk switch outlet associated with the same talking path is connected to the same control circuit or junctor. The size of the switches on each type of frame and the number of frames in each particular office will be determined by the number of subscribers and other offices connecting to this office and the calling habits of the subscribers served.

The operation of the switching network may be explained by reference to Fig. 10 which shows the control lead diagram of a skeletonized switching network of a large size office. This figure shows two line frames, each of which has two primary switches and two secondary switches. Three vertical inlets are provided on each primary switch and two vertical outlets on each secondary switch. The figure also shows two trunk frames, each of which has two primary switches and two secondary switches. The trunk switches provide two vertical trunk inlets on the primary switches and two vertical outlets on the secondary switches. Eight junctors are required as indicated. This switching network then serves to interconnect twelve subscribers with eight trunk appearances. This is the actual size built in the experimental model.

As shown in Fig. 10 each control lead path between a primary and a secondary switch on both the line and trunk frames is connected through a high value of resistance to a -45 -volt power supply. In addition each control lead path from a secondary switch terminates in a similar resistor connected to a -105 -volt power supply. In a junctor involved in an established connection, such as junctor 5 of Fig. 10, the control leads connect to a -24 -volt source through low resistance relay windings. A talking path is shown as fully established between line C on line frame 2 and trunk D on trunk frame 2. This connection is held by the current flowing from the -24 -volt source in junctor 5 through the operated reed crosspoints in the line frame to a ground in the line circuit and in the same manner through the operated reed crosspoints in the trunk frame to a ground in the trunk circuit. The -24 -volt potential on the junctor leads and the resulting -12 -volt potential on the primary-to-secondary switch link leads are effective path busy indications for subsequent path selection operations in the network.

If a talking path is now desired between line A on line frame 1 in Fig. 10 and trunk B on trunk frame 2, a +80-volt power source is connected to the control leads at these points. These applied voltages are called "marks" and originate in a number group circuit. The +80-volt mark at line A in conjunction with the -45-volts supplied to the primary-secondary switch links causes the cold cathode gas diodes of the line A vertical to fire and conduct at low current. The substantially constant voltage-drop characteristic of gas diodes causes the voltage on the two horizontal outlets of this primary switch to shift to +20 volts thereby "marking" one input lead on each secondary switch of this line frame. These +20-volt marks in conjunction with the -105 volts supplied from the junctors causes the gas diodes between the marked secondary switch inlets and the junctor outlets to fire, to conduct at low current and thereby to mark the associated junctors with -40 volts again by virtue of the gas diode characteristic. As indicated by the shaded diodes in Fig. 10 a mark on line A results in marks on junctors 1, 2, 3 & 4 and thus reveals all the idle paths from line A through the line frame.

In a similar manner the +80-volt mark applied to trunk B results in the firing of the diodes along the idle paths from this trunk to junctors 2, 4 and 7. The path to junctor 5, which is in use on the connection between line C and trunk D, is not marked in this case. The -24 volts presented by junctor 5 on its trunk control lead is not sufficient when combined with the +20-volt mark on the trunk primary-secondary link which leads to this junctor to fire the associated crosspoint diode.

For this desired connection there are two possible paths, either through junctor 2 or through junctor 4, as indicated by the -40-volt marks existing on both the line and trunk sides of these junctors. Selection between these paths is automatically accomplished by use of a lockout circuit which is common to all junctors serving the same line frame.

It is known that if a conduction path through a negative resistance gas tube is provided with a load impedance of proper value which is common to a similar conduction path through one or more other similar gas tubes, only one tube will ionize and remain ionized even if firing potentials are applied to several tubes either simultaneously or in sequence. Such a circuit employing two or more gas tubes with a common load impedance functions as a lockout circuit. The phenomenon is due to the region of negative resistance in the characteristics of the gas tube through which the tube current passes in the range between the breakdown and sustaining voltages. In this region as the current through a tube increases, the voltage across the tube decreases, tending to prevent other tubes with the common load from firing. To reduce the possibility that two

tubes fired simultaneously will then travel through this unstable region exactly together, an inductive element is used in the common load circuit. This increases the time interval required to traverse the unstable region thereby permitting differences between tubes to result in lockout.

In each junctor a five-element cold-cathode gas tube is used for path detection and selection. One control element of this tube is marked from the line side and other control element from the trunk side of the junctor if this junctor is usable in the call being set-up. The main anode is connected, together with those of the other juncctors of the same line frame, in a lockout circuit so that only the gas tube in one junctor can conduct in its main gap. The junctor in which the gas tube does conduct in the main gap is the selected junctor and the switching network path associated with it is the selected path. Assume that junctor 2 is so selected. It first shorts out the resistors in its -105 -volt supply leads. This permits a higher value of current to flow through the gas diodes along the selected path and causes the operation of the reed contacts associated with the crosspoint relay windings which are in series with the diodes. The control lead contact at each of these crosspoints, as shown along the selected path in Fig. 10, shorts out the gas diodes. With the diodes shorted out a further increase in the current operates relays in series with this control lead path in the line and junctor circuits. These relays cause the -105 -volt supplies in the associated junctor, junctor 2 in this case, to be replaced by the -24 -volt sources and the $+80$ -volt marks on the line and trunk terminals to be replaced by ground. This shift of power sources permits the gas diodes along paths marked but not selected for this call to extinguish but holds at a low power level the crosspoint relays along the selected path. With all diodes extinguished the switching network is ready for the next path selection operation. Removal of the ground at the trunk end of an established connection, at the end of conversation, results in complete release of the associated operated crosspoints and junctor.

With a central office traffic rate during busy hours of 50,000 calls per hour, 50 milliseconds is the maximum allowable holding time for a single common control circuit at 70 per cent usage. A single control circuit, even during its busiest periods, should not be in use more than about 70 per cent of the time. If the usage is increased beyond this point the delays which other circuits encounter in attempting to use the common control circuit increase very rapidly. This produces the same effect as increased control circuit holding time.

The holding time of the control circuit for the switching network determines the traffic capacity of the switching arrangement if only a

single control circuit is provided. The control circuit holding time, in turn, consists of three parts: operate and release times of connector relays, line testing and "marking" times, and the operate time of the switches and junctors. The average holding time for the control circuit of the switching network for the system described was about 40 milliseconds. This is considerably shorter than the maximum 50 milliseconds permissible under the heavy traffic conditions of the preceding paragraph.

SYSTEM OPERATION

An experimental skeletonized ECASS constructed for laboratory tests is shown in Fig. 11. The equipment is located on these frames from left to right as follows: Frame No. 1, line and originating actuator circuits, switching network and controls; Frame No. 2, trunk, outward actuator and number group circuits; Frame No. 3, originating receiver circuits; Frame No. 4, power supplies; and Frame No. 5, terminating receiver circuits.

Without further detailed description of the various component circuits the successful placing of a call through the system may now be traced by reference to the block diagram of Fig. 12.

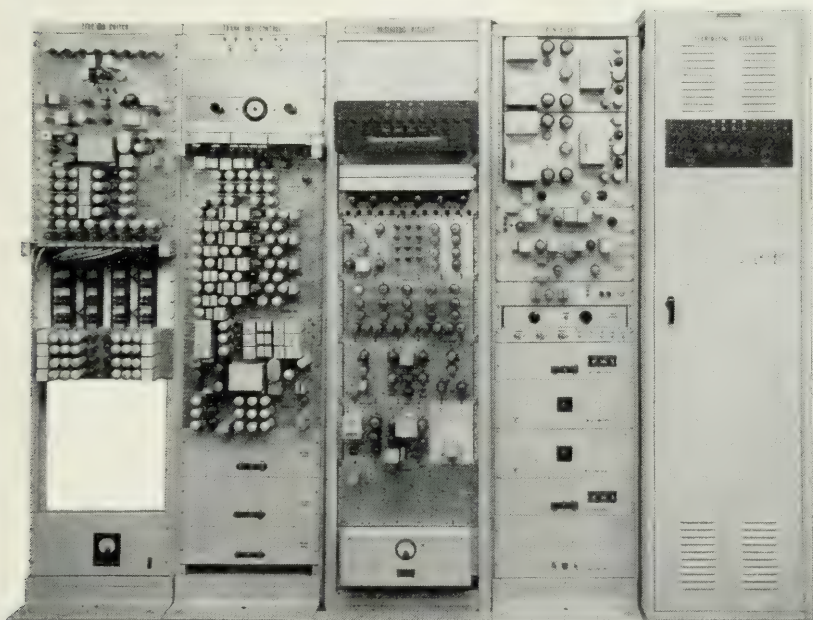


Fig. 11—Skeletonized laboratory model of ECASS.

A subscriber in originating a call first pre-sets the complete called line number on the finger wheels of his subset. The subset has been "latched" in the signaling condition by the mechanical reset on hanging up after the previous call. When the subscriber then removes the handset of his telephone from its cradle a line relay in the central office operates in recognition of a demand for service. The line relay in turn energizes a start gap of an associated cold-cathod gas tube. The gas tubes for a group of lines are connected in a lockout arrangement such that only one gas tube at a time can conduct in a main gap. When the tube does conduct in the main gap it operates a relay which connects the associated line directly to a common originating actuator and receiver circuit. During the short period that one line is attached to the receiver, originating service is withheld from all other lines in the same group but incoming calls may be terminated to any idle line.

The name, actuator, in this system refers to a circuit which includes an amplifier for transmitting 50-cycle current to a subscriber's subset over the simplex. This current is maintained at a constant amplitude despite the differences between various subscriber loops and the possible presence of earth potentials by the high output impedance of the amplifier. This high output impedance is obtained by the use of 35 db of feedback from the output of the amplifier. In addition, the actuator circuit also monitors its 50-cycle current flow when connected to a subscriber loop as the means of maintaining supervision since no direct current is permitted in the loop during the signaling period. The 50-cycle current in the subscriber's set causes the complete pre-set number to be generated repetitively as pulse position dialing signals which are returned to the receiver circuit in the central office over the loop. The use of simplex power to generate loop signals was adopted to simplify the filtering problem at the receiver circuits.

The originating receiver detects the dialing signals including the occurrence of the blank interval between repetitions of a complete number. It decodes the signals representing the first three digits following the blank interval, i.e., the called office code, and registers these digits unless the check circuit indicates that another trial is necessary. The action of the check circuit has been described in the Signal Receiver section of this paper. The receiver ignores the signals representing the called line number. Upon the successful registration of the called office code the originating receiver connects to the trunk number group circuit.

The name, number group circuit, in this system refers to a circuit through which a connection may be made to the switching network appearance of the control lead of any of a group of trunks or lines. In

the trunk number group a matrix of cold cathode gas tubes combines the three digits of an office code to establish a single lead control path to the equipment appearances of the trunks. This translator feature permits an arbitrary assignment between trunk locations and directory listed office codes. Another circuit, the subscriber number group, similarly includes translation of a called line directory number into the switching network line equipment number. Over such a control path a test is made of the idle, busy, or vacant condition of any designated trunk or line, and this same control path is used, together with other control leads to the switching network, to establish a connection through the switching network to this trunk or line.

If the test through the number group discloses an idle trunk, the control terminal of the trunk appearance on the reed-diode switches is "marked" with voltage over the same busy-testing path and the control lead of the calling line appearance is similarly "marked" over a path extending through the receiver-actuator connector. These marks from opposite ends of the switching network cause the selection of an idle junctor located in the connecting leads between line and trunk frames. The selected junctor in turn functions to make the marks effective in operating the switch crosspoints of all four switch stages as described

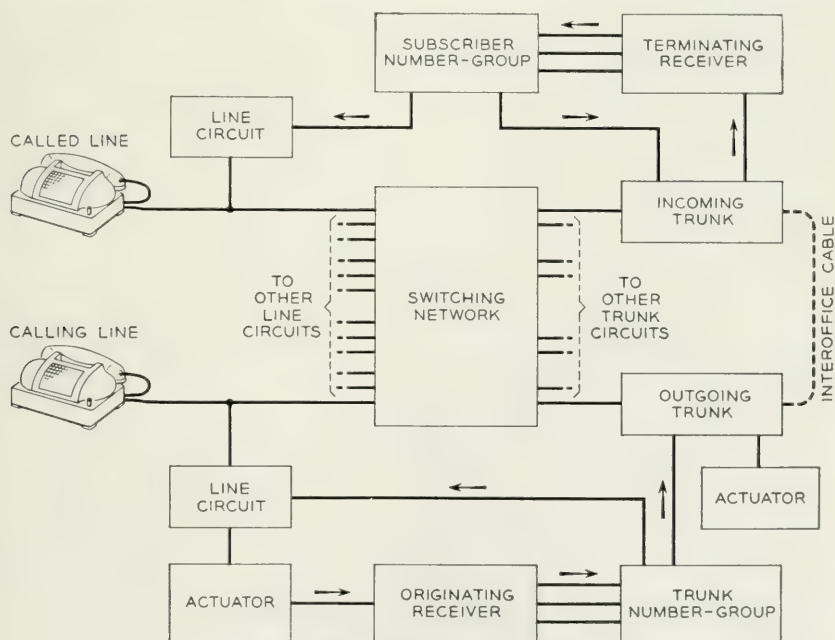


Fig. 12—Block diagram of ECASS.

in the section on Switching Network. The marking and switch operating voltage is applied to the line terminal of the switches through a line cut-off relay, which operates on the increased current which flows in this circuit immediately after the diodes in the switch crosspoints have been shorted out. The operation of the line cut-off relay releases the originating actuator and receiver which were connected through back contacts of this relay and, in turn causes the release of the trunk number group.

The next step is to send the called line number over the outgoing trunk so that the distant office may complete the connection to the called subscriber. An outward actuator is provided for this purpose. A relay in series with the marking path in the outgoing trunk circuit operates to connect the outward actuator directly to the trunk. The trunk-to-actuator connector circuits include gas tube lockout to insure that only one trunk is connected to the actuator at a time. During the short delay of awaiting an actuator that may occur during heavy traffic periods the established switching network connection is held under control of direct current supervision from the trunk circuit. The outward actuator, when connected, transmits 50-cycle current through the switching network to the calling subscriber's subset and maintains the connection by monitoring the 50-cycle current flow. This 50-cycle current causes the subscriber's set to transmit again the called number repetitively through the switches and outgoing trunk to the associated incoming trunk at the called office. In this paper it is assumed that all other offices connecting to this one are of the same type as this one or are arranged to transmit and receive, when required, the signaling pulse code used in this office. The arrangements in this office for completing incoming calls, including calls originating within this office itself, are shown in Fig. 12.

Operation of the connector relay which connects the outgoing trunk to an actuator signals the incoming trunk circuit in the terminating office to connect to an incoming receiver circuit for receiving the repetitive dialing signals. Connection between the incoming trunk and signal receiver is made through a lockout circuit which insures that only one trunk is connected to the receiver. When the incoming receiver has absorbed the office code, registered the called line number and checked the registration, it causes the incoming trunk to transmit a reverse battery pulse to the outgoing trunk as a number-received signal. This reversal causes the outgoing trunk to dismiss the outward actuator and to trip the latch in the subscriber's subset to the talking position with direct current talking and supervisory battery supplied from the outgoing trunk. At the same time the incoming receiver connects to the subscriber number group for making an idle-busy-vacant test of

TABLE I

Connecting Times	Milliseconds
1. Calling line off-hook to connection to outgoing trunk.....	180
2. Incoming trunk seizure to ringing of called line.....	200
Total time to establish a call.....	380
Holding Times	Milliseconds
1. Originating receiver.....	165
2. Trunk number group.....	38
3. Switching network control (each usage).....	14
4. Outward actuator.....	291
5. Terminating receiver.....	184
6. Subscriber number group.....	38

the called line and for "marking", if idle, the called line control terminal appearance on the reed-diode switches. At the time of this test, a voltage "mark" is applied to the incoming trunk control lead appearance also. As before, these two "marks" from opposite ends of the switching network cause the selection of an idle junctor and in turn the operation of the reed crosspoints in the four switching stages along the selected path. The "marking" voltage is applied to the incoming trunk terminal of the switches through the winding of a relay which, operating immediately after the crosspoints, causes the release of the incoming receiver and places the switching connection under joint supervision of the called and calling subscribers. The line cut-off relay whose winding is in series

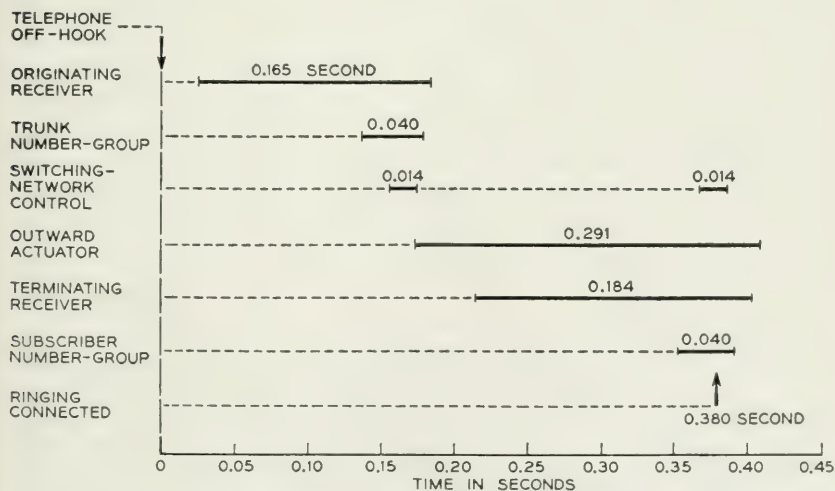


Fig. 13—Operating sequence based on average holding times.

with the marking path to the line appearance on the switches operates to remove the line relay and other originating apparatus from the called subscriber's line.

Based on the result of the idle-busy-vacant test of the called line, the incoming register circuit either sets the incoming trunk to provide ringing to the called subscriber upon closure of the crosspoints and ringing tone to the calling subscriber if the called line is idle, or sets the trunk to return busy tone to the calling subscriber if the called line is busy or vacant. In the latter case the incoming receiver is released immediately without setting up a terminating connection through the switches.

Since connections in the same reed-diode switching network are established through either of two number group circuits, lockout is provided between the originating receiver to trunk number group connector and the terminating receiver to subscriber number group connector so that only one number group circuit can be in operation at a time.

Some of the important average time intervals as measured in this system are given in Table I and shown graphically in Fig. 13.

CONCLUSION

The electronically controlled automatic switching system described in this paper was designed for large central offices and a skeletonized laboratory version has been built, tested and demonstrated. Successful operation at the speeds required was obtained. No failures of the gas tube lockout circuits were observed under the various combinations of possible simultaneous seizure. The experimental system shows that a large heavy traffic office could be made to operate on a one-at-a-time basis with advantageous reduction in the number of control and connector circuits. Many of the necessary components employed in this system for one-at-a-time operation are now available in a pre-development state and will probably be used in commercial systems. However, the commercial design and production of a complete office as described here is not economically competitive with existing systems since the subscriber subset and line circuit which are used in large numbers are too complex and expensive.

ACKNOWLEDGEMENTS

Although acknowledgements have been made in specific cases throughout this paper, we wish to point out that many others contributed to the success of the project. We wish to mention G. G. Bailey and G. A. Backman who performed the physical construction and assisted in the testing. In particular we wish to mention A. W. Horton, Jr., who directed the project.

New Techniques for Measuring Forces and Wear in Telephone Switching Apparatus

By WARREN P. MASON AND SAMUEL D. WHITE

(Manuscript received February 15, 1952)

One of the main problems in obtaining long life in telephone switching equipment is the wear caused by large momentary forces. In order to investigate this problem several new techniques have been devised for measuring normal and tangential forces and for producing and controlling normal and tangential motions for wear studies. The forces are measured by inserting small barium titanate ceramics between the points of application of the forces and observing the voltages generated on a cathode ray oscillograph. Barium titanate ceramic is about fifty times as sensitive as quartz and has a high enough dielectric constant so that with conventional amplifiers time intervals as long as a tenth second can be measured. Both normal and tangential forces can be measured by using properly poled ceramics. By using weights on top of the crystals, normal and tangential accelerations can be measured. With these ceramics, forces have been measured for relays and for frictional sliding of a wire over a plastic. By employing a barium titanate transducer capable of a large amplitude at 18,000 cycles it has been shown that no wear occurs for normal forces, and that all the wear observed in a relay is due to tangential sliding. Quantitative measurements of wear have been made for a variety of materials, and it has been shown that materials with a large elastic strain limit will wear better than materials with a small elastic strain limit even though the latter have a higher yield stress; materials such as plastics and rubber will outwear materials such as metals or glasses.

As the length of slide is reduced there is a threshold of motion for which there is no gross slide and very little wear. This region is determined by the condition that the tangential force is smaller than the normal force times the coefficient of friction. Theoretical and experimental results are obtained for this region and an equation is derived which determines the possible displacement without gross slide. The stress strain curve occurs in the

form of a hysteresis loop whose area varies approximately in proportion to the square of the strain amplitude. This region is important for relays for by introducing damping, long repeated vibrations—which are responsible for considerable wear—are quickly brought down to the low wear, no gross slide region with a corresponding reduction in wear. The mechanical resistance associated with the stress strain loop is of the same type that occurs in an assemblage of granular particles such as in a telephone transmitter where the motion is small enough so that no gross slide occurs.

I. INTRODUCTION

In obtaining long life in telephone equipment such as relays, switches, selectors and other mechanical devices subject to large momentary forces, one of the main problems is the wear encountered in various parts. This is particularly true in such small motion devices as relays where even a few mil inches of wear increases the distance that the armature has to travel and may eventually cause the relay to fail to make contact. To obtain a design objective of one billion operations requires a very careful minimizing of deleterious forces and a careful selection of the best wearing materials.

As a step toward investigating this problem several new techniques have been devised for measuring normal and tangential forces and for producing and controlling normal and tangential motions for wear studies. These methods have been applied to relays and have given considerable information on the types of motion to be avoided and on the best types of materials to select for various parts of the relay to obtain long life. Specifically they have shown that normal forces cause very little wear and that tangential sliding of one part over another is the principal cause of wear. Fortunately, by designing the motion of the armature and contacts correctly, tangential sliding can be largely eliminated with a corresponding reduction in wear.

To aid in the quantitative evaluation of wear produced by tangential sliding two devices have been used. One is an electromechanical vibrator¹ driven at 500 cycles per second which is capable of several mil inches of motion and the other is a barium titanate longitudinal vibrator coupled to a metal "horn"² which is capable of a two mil inch motion at 18,000 cycles. Wires connected to these transducers are dragged over materials whose wearing properties are to be tested. The normal forces between the wire and material are varied as well as the length of the stroke. The wear by both methods is comparable showing that the accelerated wear testing method gives about the same wear as the slower

method. With the barium titanate transducer a billion cycles can be obtained in 17 hours and a very rapid wear test is obtained.

If lubrication is not used, wear tests show that materials having a large elastic strain limit will in general wear better than materials which have a smaller elastic strain limit even though the latter may have a higher yield stress; materials such as plastics and rubbers will outwear materials such as metals and glasses. The volume of wear for one billion operations is proportional to the product of static force times the length of the stroke. The initial rate of wear is several times as large as the final rate. Calculations show that only about one part in 10^9 of the energy goes into producing wear, the rest going into heat production.

As the length of slide is reduced, calculations and measurements show that there is a threshold of motion for which no gross slide occurs. This condition occurs when the tangential force is less than the product of the normal force times the coefficient of friction. The limiting displacement for no slide increases as the two-thirds power of the normal load and inversely as the two-thirds power of the shear stiffness. Hence a heavily loaded material with a small shear elastic constant—such as rubber—will have a large displacement for which no slide occurs, and hence will wear considerably better than a stiff material such as a metal. Wear tests in the region of no gross slide show that the rate of wear is considerably less in proportion to the energy dissipated than in regions of gross slide.

A quantitative experimental and theoretical study of the region of no gross slide has been made.³ Experimentally the results have been obtained by moving a glass lens with a large radius of curvature on both surfaces between two glass lenses when the lenses are pressed together with known normal forces. It was shown theoretically that slip should occur between these lenses over a circular annulus and experiments verify this prediction quantitatively. Force-displacement curves have been measured and it has been shown that the relation is a hysteresis type loop whose area varies approximately as the square of the strain amplitude. The small wear observed is related to the wear found in ball bearings, where no gross slide occurs. This region is important in relays for by introducing damping, long repeated vibrations—which are responsible for considerable wear—are quickly brought down to the low wear, no gross slide region with a corresponding reduction in wear. The mechanical resistance associated with the stress strain hysteresis curve is of the same type that occurs in an assemblage of granular particles such as in a telephone transmitter, where the motion is small enough so that no gross slide occurs.

II. METHODS FOR MEASURING NORMAL AND TANGENTIAL FORCES

In order to investigate the performance of a mechanical device and the causes of wear in it, it is desirable to be able to measure the forces occurring in various parts of the device. To measure the complete performance it is necessary to measure not only the slowly applied forces but also the very short time dynamic forces that occur when various parts of the device impinge on each other.

The most common method for measuring such forces is by means of a piezoelectric crystal such as quartz. Quartz, however, has the disadvantage that it is not very sensitive and also that it has such a low dielectric constant that the input impedance of any amplifier associated with it has to be prohibitively high if forces varying as slowly as a one-tenth of a second are to be measured. Since the impedance of the oscillograph or amplifier is usually lower than that of the crystal, the crystal having the greatest sensitivity will be the one which generates the most charge for a given force, which corresponds to the crystal having the largest d piezoelectric constant. Table I shows a tabulation of the d constants for compression and shear for several of the most common piezoelectric crystals and for the ceramic barium titanate. The dielectric constants are also given.

Of these materials the only ones that have sufficient mechanical strength to withstand the mechanical shocks they are subjected to in the measurements of forces are quartz, tourmaline and barium titanate ceramic. The crushing strength of the ceramic has been found⁴ to be from 60,000 to 80,000 pounds per square inch. From the values of the d piezoelectric constants it is seen that the barium titanate ceramic is about 50 times as sensitive as quartz or tourmaline and it is possible to use small pieces of the ceramic to work directly into cathode ray oscillo-

TABLE I

Crystal	Compression Constant in cgs units	Shear Constant in cgs units	Dielectric Constant
Quartz.....	$d_{11} = 6.76 \times 10^{-8}$ stat. coulombs/dyne	$d_{26} = 13.5 \times 10^{-8}$ stat. coulombs/dyne	4.55
Tourmaline..	$d_{33} = 5.5 \times 10^{-8}$	$d_{16} = 10.9 \times 10^{-8}$	8.0
ADP.....	$d_{31}' = 74 \times 10^{-8}$	$d_{36} = 148 \times 10^{-8}$	15.6
Rochelle Salt			
Y Cut.....	$d_{21}' = 84.5 \times 10^{-8}$	$d_{26} = 169 \times 10^{-8}$	11.1
EDT.....	$d_{21} = 34 \times 10^{-8}$	$d_{36} = 50 \times 10^{-8}$	8.0
Barium Titanate Ceramic.....	$d_{33} = 300 \text{ to } 400 \times 10^{-8}$	$d_{36} = 500 \text{ to } 650 \times 10^{-8}$	900 to 1500

graphs with the use of only the amplifiers that are included with such oscillographs. To work down to time intervals in the order of one-tenth second, the leakage resistance of the load across the polarized ceramic—which for small sizes may have a capacitance as low as 20-micro-microfarads—has to be higher than is usually available in oscillographs. Fig. 1 shows a vacuum tube circuit⁵ capable of giving a 750-megohm input resistance and when used with a barium titanate ceramic having a capacity of 20 $\mu\mu\text{f}$, allows measurements of forces for time intervals up to 0.015 seconds with no corrections. This time is usually sufficient to obtain all the force variations in a relay operation. The upper frequency limitation in the measurements of forces is caused by the setting up of natural vibrations in the ceramic block. The lowest frequency vibrations that can be set up in a ceramic block are the flexural vibrations. For a block 0.04 inch x 0.04 inch in cross section and 0.02 inch thick, such as have been used in relay force measurements, the lowest flexural frequencies are in the order of 1.6×10^6 cycles. The next lowest frequencies are the radial mode vibrations⁶ which have frequencies above 4 megacycles for the block considered. Hence the measurements of force should be valid up to times in the order of a microsecond.

The properties of barium titanate and their stability with time and with large voltages applied in the opposite direction to the poling voltage depend to a large extent on the method of baking the ceramic and on

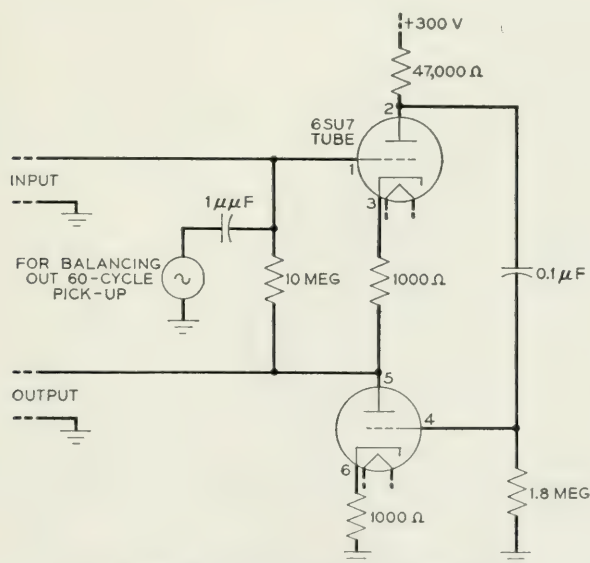


Fig. 1—High input resistance amplifier tube.

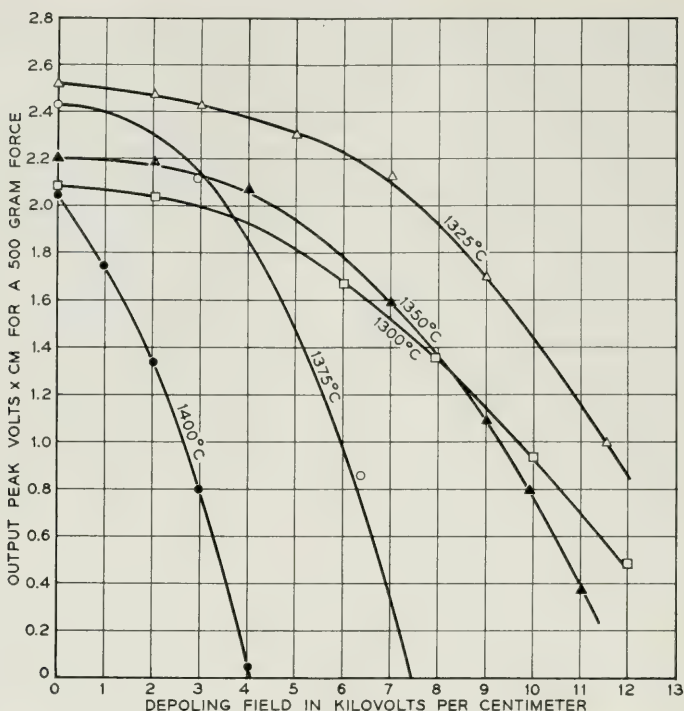


Fig. 2—Open circuit voltage for 500 grams force for normal barium titanate under depoling voltages.

the effect of additives. The data of Fig. 2 show⁷ the effect of firing temperature on the initial open circuit voltage for sample disks 0.775 cm in diameter and approximately 0.15 cm thick. The variation of voltage with thickness and area was taken account of by multiplying the measured voltage by the area and dividing by the thickness.

The open circuit voltage was measured by using the circuit of Fig. 3. A barium titanate cylinder and metal horn described in a previous paper,² vibrating at 18,000 cycles, strikes the sample a blow at its central position. The voltage generated is applied to the input of a high resistance tube similar to the one shown by Fig. 1, and then actuates a cathode ray tube. The voltage corresponding to the height of the peak is calibrated by putting a known voltage in series with the ceramic across a small resistance R and hence the magnitude of the open circuit voltage can be quantitatively determined. The value of the mechanical blow applied to the polarized ceramic can be adjusted by controlling the drive on the ceramic cylinder. With the feed back circuit described in the

previous paper,² this value can be held very constant and can be controlled by controlling the bias on the limiting device. The magnitude of the force can be determined by comparing the voltage with that obtained by suddenly lifting a weight off the ceramic and has been adjusted to equal 500 grams. The voltages shown then correspond to the open circuit voltages generated by applying 500 grams to a point at the center of the ceramic.

As shown in the appendix, the effect of applying a force at a point in a ceramic is not the same as that caused by distributing the force uniformly over the surface due to the fact that radial strains are generated and these act through the radial piezoelectric constant to reduce the value generated by the thickness piezoelectric constant. It is shown that the point application of stress generates only 40 per cent as much as would

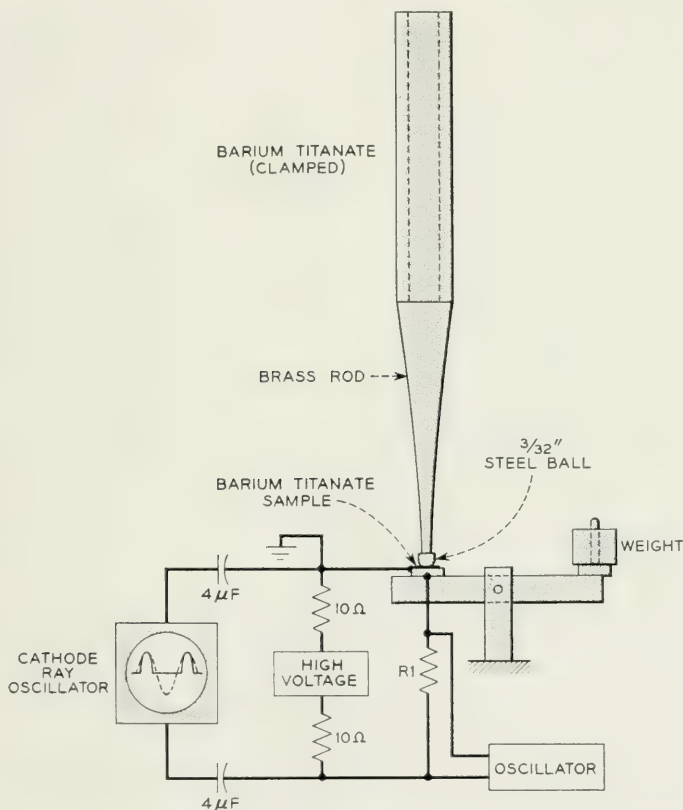


Fig. 3—Circuit used to measure open circuit voltages for barium titanate samples.

be generated in a disk with the stress applied uniformly. With this factor the open circuit voltage per unit of force—which determines the effective g_{33} piezoelectric constant of the ceramic—agrees well with that obtained by other methods of measurement. For the most desirable ceramic obtained for the 1325°C baking temperature the value of g_{33} equals

$$g_{33} = 3.25 \times 10^{-8} \frac{\text{cm}^2}{\text{stat-coulomb}} \text{ in e.g.s. units} = 0.98 \times 10^{-2} \frac{\text{meters}}{\text{Newton}} \text{ in m.k.s. units} \quad (1)$$

The dielectric constant ϵ of this material is about 1500 so that the d_{33} piezoelectric constant is

$$d_{33} = \frac{g_{33}\epsilon}{4\pi} = 390 \times 10^{-8} \frac{\text{stat-coulombs}}{\text{dyne}} = 130 \times 10^{-12} \frac{\text{coulombs}}{\text{Newton}} \quad (2)$$

Since for some applications in this paper, high voltage gradients of opposite sign to the poling voltage are applied to the ceramic, it is a matter of importance to find out whether the ceramic will become depoled by the action of this voltage. To test out this feature the circuit of Fig. 3 is equipped with a high voltage generator, which is applied to the ceramic through 10-megohm resistors and the high voltage is kept out of the measuring circuit by 4-microfarad condensers. The procedure was to apply a negative voltage for 3 minutes, then to recalibrate the voltage due to impact. This was repeated with a higher voltage each time until the range was covered.

The curves of Fig. 2 show that there is an optimum baking temperature for a large coercive field. Above this temperature larger sized crystals grow in the ceramic and the coercive field decreases markedly. It is thought that the smaller crystal size corresponds to a more strained condition in the individual crystallites and it requires a higher field to overcome the mechanical bias and change the direction of the ferroelectric axis. A similar condition⁸ has been found by x-ray techniques for single crystals where it has been found impossible to make a single domain out of a multidomain crystal by the application of a field, if the crystal is too highly strained.

The effects of additives are also very marked on the properties of the polarized ceramics. It has previously been reported⁹ that the addition of 4 per cent of lead titanate to the commercial barium titanate increases the coercive field. This is confirmed by the curves of Fig. 4 which show

the open circuit voltage for a 4 per cent lead titanate barium titanate ceramic for various baking temperatures and negative biasing voltages. The optimum temperature for a small grain size structure is lowered about 50°C by the addition of the lead titanate. As can be seen the coercive field is considerably increased and it appears safe to use a negative field of 6000 volts per centimeter without any depolarization. In Section IV a system is described for which an ac voltage of this magnitude was successfully used for many days with no change in sensitivity of the ceramic. The open circuit piezoelectric constant for the optimum ceramic of Fig. 4 is

$$g_{33} = 3.82 \times 10^{-8} \frac{\text{stat-coulombs}}{\text{dyne}} = 1.15 \times 10^{-2} \frac{\text{meters}}{\text{Newton}} \quad (3)$$

Since the dielectric constant is about 1000, the effective d_{33} piezoelectric constant is about

$$d_{33} = 310 \times 10^{-8} \frac{\text{stat coulombs}}{\text{dyne}} = 104 \times 10^{-12} \frac{\text{coulombs}}{\text{Newton}} \quad (4)$$

Another property of interest is the stability of the piezoelectric properties of the ceramic over a long period of time. While no very good comparisons have been made between the various baking conditions and between barium titanate with and without additions, some long time measurements have been made on four samples of the optimum 4 per cent lead titanate used in the transducer of Fig. 3. Over a period of two years during which they have been continuously used in a calibrated oscillator, the calibration has not changed noticeably, i.e. less than 5 per cent. On account of the superior voltage and time stability of the lead titanate, barium titanate mixture, all of the elements used have had this composition.

Two types of units have been used for force measurements, one type that responds to normal forces and the other to tangential forces. The type responding to normal forces as shown by Fig. 5 is poled in the thickness direction which is also the direction in which the force is applied. The sensitivities for forces applied at points are given by the values of Fig. 4. For example for typical units having the dimensions 0.1 cm by 0.1 cm in cross section and 0.05-cm thick will produce an open circuit voltage of 2.7 volts for 100 grams applied to the ceramic. Such ceramics have been used in measuring the dynamic forces when various parts of the relay close or open. Fig. 6(a)¹⁰ shows the voltage generated when the two relay contacts come together. The dynamic stress is somewhat higher than the static stress and varies with time due to mechanical

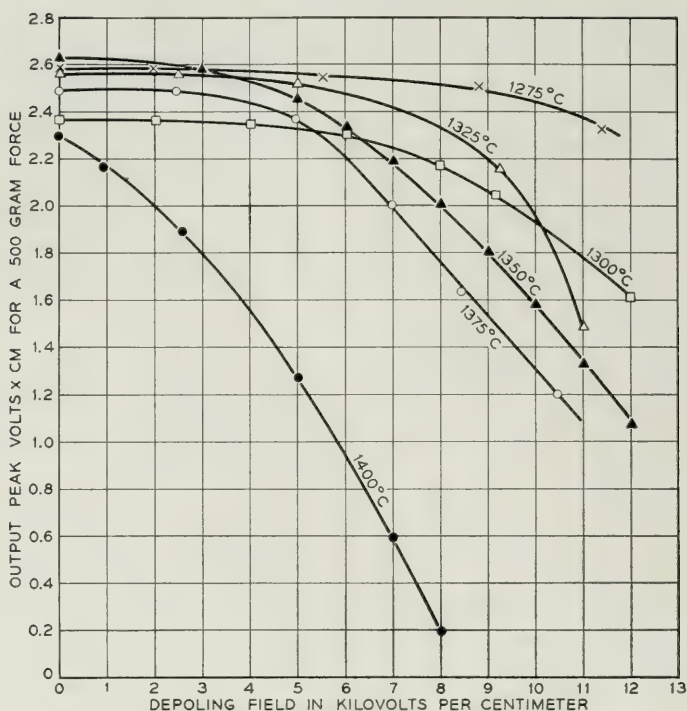


Fig. 4—Effect of 4 per cent lead titanate on the open circuit voltages generated for 500 grams force, and for depoling voltages.

vibrations of the relay structure. Fig. 6(b) shows the forces produced by opening the contacts. The large spikes are due to wire vibrations. By using such ceramics in various parts of the relay the points of high stress can be located.

The second type of structure which responds to tangential forces is poled as shown by Fig. 5 so that the poling direction lies along the direction for which the tangential force is applied and perpendicular to the

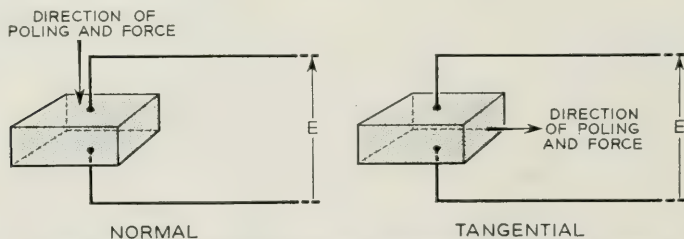


Fig. 5—Methods for polarizing barium titanate to respond to normal and tangential forces.

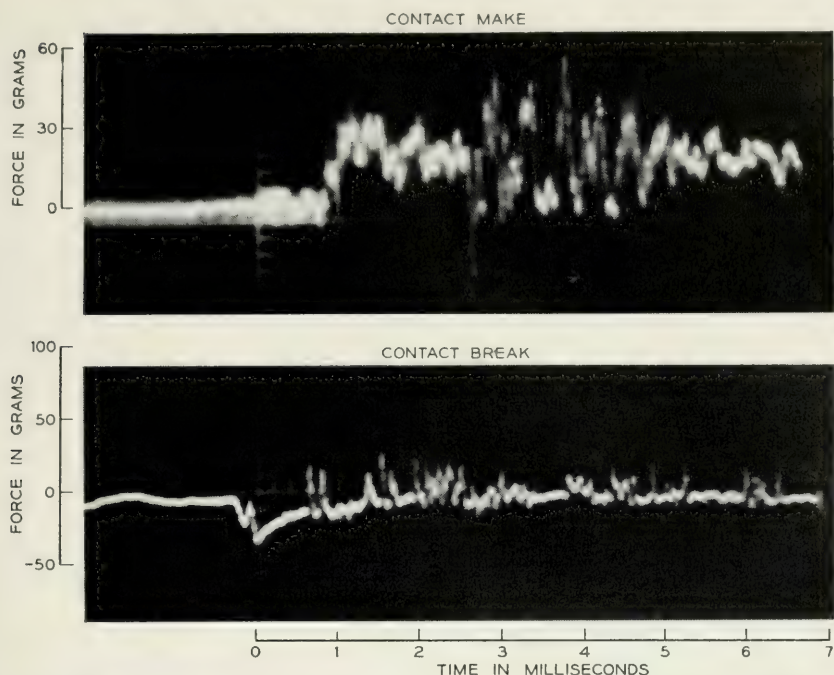


Fig. 6—Oscillograph tracings of forces generated in make and break operations.

direction of the electrodes. In this process, the crystal is first poled, after which the poling electrodes are ground or etched off and electrodes perpendicular to the poling direction are put on by using a polymerizing cement in which silver dust is mixed. The cement serves not only as an electrode but also holds the ceramic in the desired place. Fig. 7 shows an arrangement used for studying frictional forces. A small ceramic 0.1 by 0.1 cm in cross-sectional area is glued to a metal base while a thin specimen of the material whose frictional forces are to be studied is glued to the top surface. The forces caused by a wire drawn over the surface are transmitted to the crystal and generate a voltage which appears on the oscillograph. Pictures of such force generated voltages are

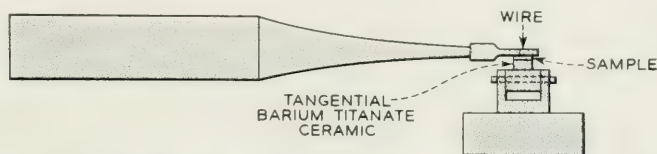


Fig. 7—Experimental arrangement for studying frictional forces.

shown by Fig. 9 of the next section and are discussed there. The sensitivity of this type of unit is higher than that for the normal force measuring unit. As shown in the appendix, the voltage generated is independent of the area of application and is about 9.7 volts for a unit the same size as discussed above which gave 2.7 volts for 100 grams applied at a point.

By placing weights on the upper surfaces both types of units can be used as accelerameters. They are cemented to the surface whose acceleration is to be measured and the force applied is equal to half the mass of the ceramic plus the added mass times the acceleration. By putting weights on the shear pickup ceramic types, tangential accelerations can be measured in the direction of the poling. By using three such accelerameters, the normal and two tangential components of acceleration of any surface can be measured.

III. METHODS FOR INVESTIGATING CAUSES OF WEAR

Wear in various parts of a relay is the limiting factor when a very large number of relay operations are desired. This wear opens up the spacing between contacts and causes the relay to lose its adjustment over a course of time.

A. Force Measurements and Wear Caused by Normal Forces

Since the forces operating on a material can be divided into normal and tangential forces, it appears desirable to separately determine the effects of each. Normal forces were produced by using the barium titanate, metal horn detail of Fig. 3. With a steel ball on the end of the metal horn, and a barium titanate specimen glued to the pivoted arm, the peak forces in grams are plotted against the volts used to drive the titanate unit for various static forces in Fig. 8. The pattern of voltage is approximately a rectified sine wave, since the ball is out of contact with the measuring titanate a part of each cycle. To observe the wear caused by normal forces a piece of material to be studied was glued to the pivoted arm on top of the barium titanate and the force was adjusted to the required value. For forces in the order of those measured in relays no wear at all was observed over a period of 18 hours which corresponds to a billion impacts, since the number per second is 18,000. For larger impulsive forces, it was found that the result of 60-million impacts against an insulator such as a phenolic was to produce a pit only a few tenths of a mil inch deep by a plastic flow. Since no wear of the type involved in relays was observed it was concluded that practically all of the wear was produced by tangential forces.

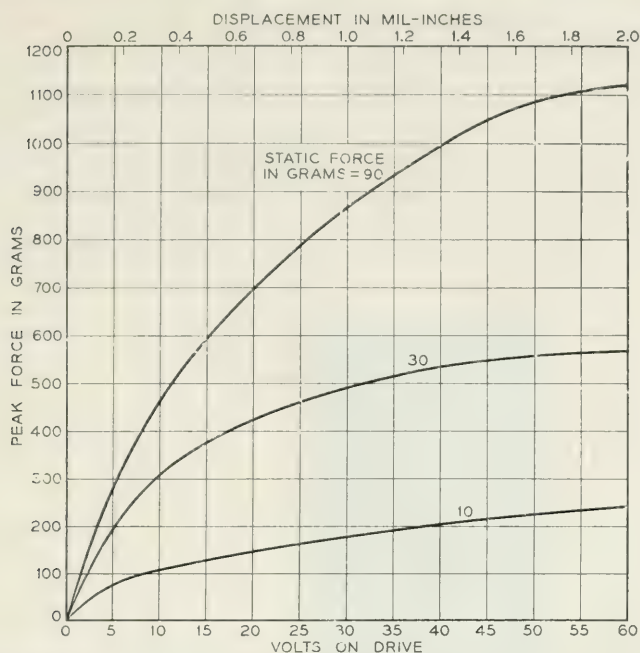


Fig. 8—Variation of normal impulse forces with drive voltages for three values of static force.

B. Tangential Force and Wear Measurements

To study the effect of tangential forces in producing wear, the transducer was mounted horizontally and the steel ball was replaced by a wire such as are used in some relays. The length of the wire was made short enough so that no lateral vibrations were generated and the motion was strictly tangential. Samples to be studied as shown by Fig. 7 were mounted on top of shear type ceramics which were glued to the pivoted arm in such a way that they responded to tangential forces applied perpendicular to the arm.

When a piece of A phenolic (which is a paper filled phenolic) was placed on top of the ceramic a series of oscillograph pictures were taken when the total displacement of the wire varied from 0.05 mil inch to 2.0 mil inches and the steady weight on the wire was 40 grams (0.0885 pound). These pictures are shown in Fig. 9. For amplitudes under 0.075 mil inch, the force is a good sine wave which increases with amplitude until the maximum force equals the product of normal force times the coefficient of friction. The force in this region is essentially elastic as is shown by the fact that the maximum force occurs at the time when

the maximum displacement of the wire takes place. Above this amplitude the wire begins to slip on the plastic and for a travel of 0.3 mil inch there are indications that the point of contact between the wire and the plastic has changed from one position to another. This agrees with the idea that friction is due to a definite bonding between points of contact of the two materials which is broken by their relative motion. New points of contacts are then made and a stick-slip process occurs. At 0.5-mil-inch motion a number of small contacts occur during the

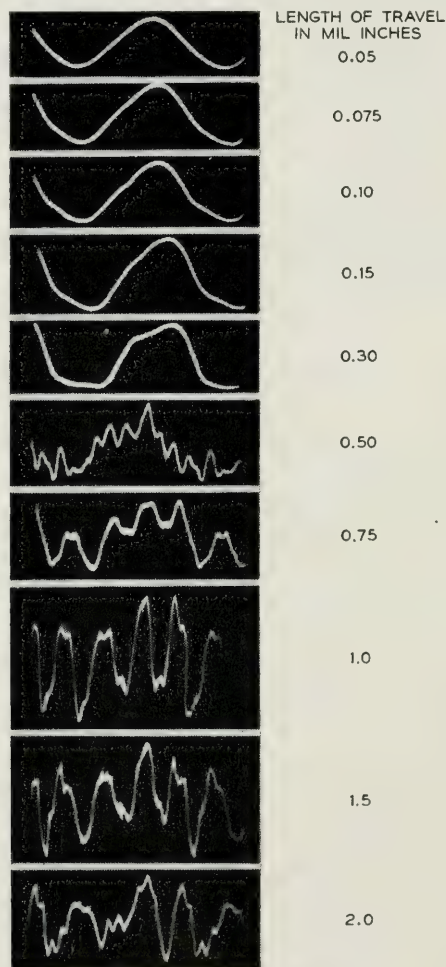


Fig. 9—Tangential forces measured for an 18,000 cycle oscillatory motion whose total displacements in mil inches are shown by the values given.

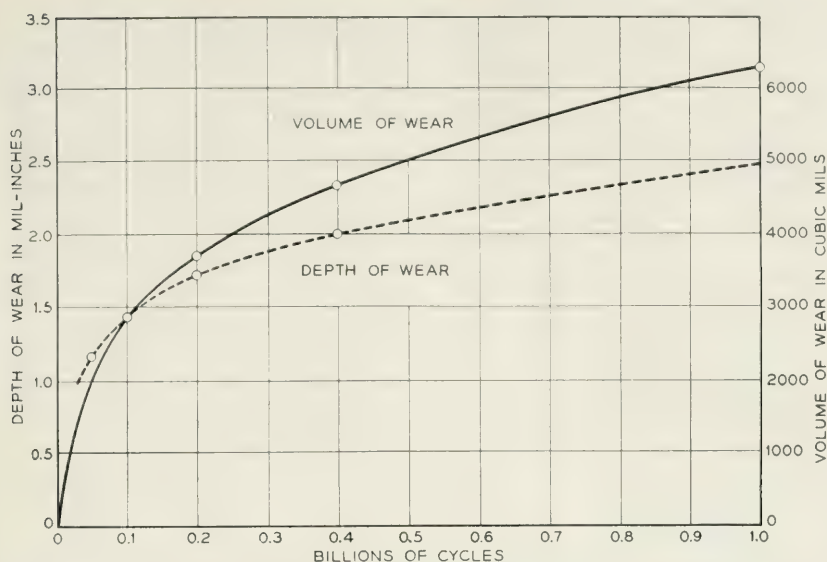


Fig. 10—Typical wear curve for A phenol fibre plotted as a function of the number of cycles.

travel. Since the picture is a trace of an oscillograph pattern which is being repeated 18,000 times a second and since a two second exposure is required to produce the picture it is obvious that the wire goes back and forth over the same points for a large number of times. Most of the energy is lost in producing elastic vibrations in the points of contact. These oscillations are produced by the bending of the areas of contact by the bonding force between them and by the motion. When the bond is broken the plastic forming the point is free to vibrate and the elastic energy goes into mechanical vibrations and eventually into heat. Since a pattern such as that for the 0.5-mil inch or the 0.75-mil inch displacement lasts unchanged for a number of minutes, it is obvious that very little of the energy goes into breaking the plastic points of contact and producing wear. This is confirmed by a rough calculation given later which shows that only about 1 part in 10^9 of the energy goes into producing wear. For displacements above a mil-inch motion it appears that groups of point contacts are broken at one time, and the pattern changes rather rapidly indicating that there is more wear at these amplitudes. Over a two-second interval the pattern is changing fast enough so that sharp pictures are not obtained.

Quantitative values of wear for various materials were obtained by running the barium titanate unit for various periods of time, different lengths of strokes and different normal forces. Fig. 10 shows a typical

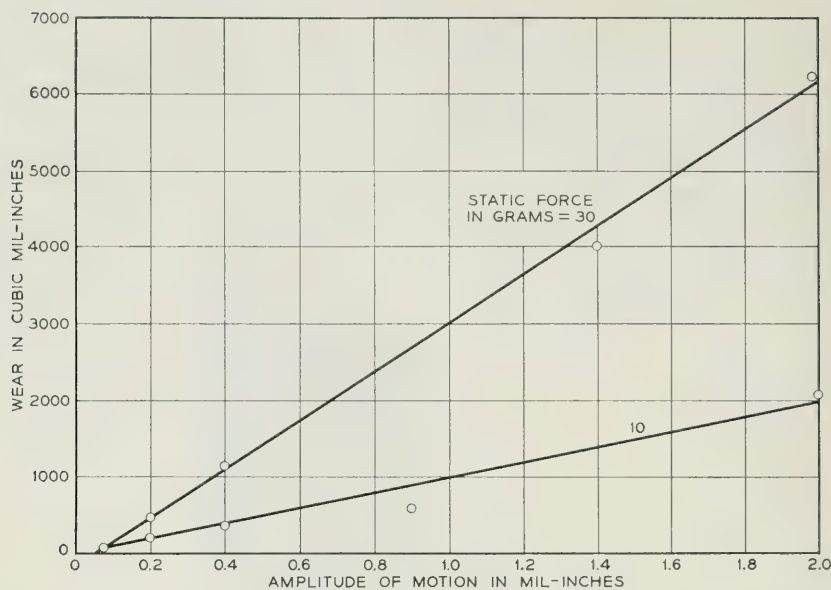


Fig. 11—Total wear for one billion cycles plotted against the length of stroke for two normal loads.

wear curve obtained for A phenolic (a paper filled phenolic) plotted as a function of the number of cycles. This wear was obtained by drawing a 0.025-inch nickel silver wire for a distance of 2.0 mil inches over the surface of the bar. The bar was $\frac{1}{4}$ inch wide. The normal force used was 30 grams (0.0665 pound). The wear was measured from the depth cut in the material and from this since the wire was round, the total volume of wear in cubic mil inches could be calculated. The rate of wear was faster at the start but approached a limiting rate with a large number of cycles.

A number of different lengths of stroke were employed and for the A phenolic the total wear for a billion operations is shown plotted by Fig. 11. The wear is approximately proportional to the slide but extrapolating down to small motions it appears that there is a threshold of motion below which the wear is very small. The values indicated are close to the no gross slide regions found from the force curves of Fig. 9 for both forces shown in Fig. 11. To check that the wear was definitely less in the no gross slide region an amplitude of motion of 0.075 mil inch for a normal force of 50 grams (0.11 pound) was run for a billion operations. The wear observed was so small that it could not be measured quantitatively, confirming the lower rate of wear in the elastic region.

Another type of wear measurement has also been employed. As shown by Fig. 12 the motor is a modification of the Western Electric 1A recorder, which was originally designed for cutting "hill and dale" phonograph records.¹ The moving system of this recorder consists of two coils (a drive coil and feedback coil) and a stylus, all rigidly coupled and coaxial. The drive coil is secured to the base of a cone shaped vibrating

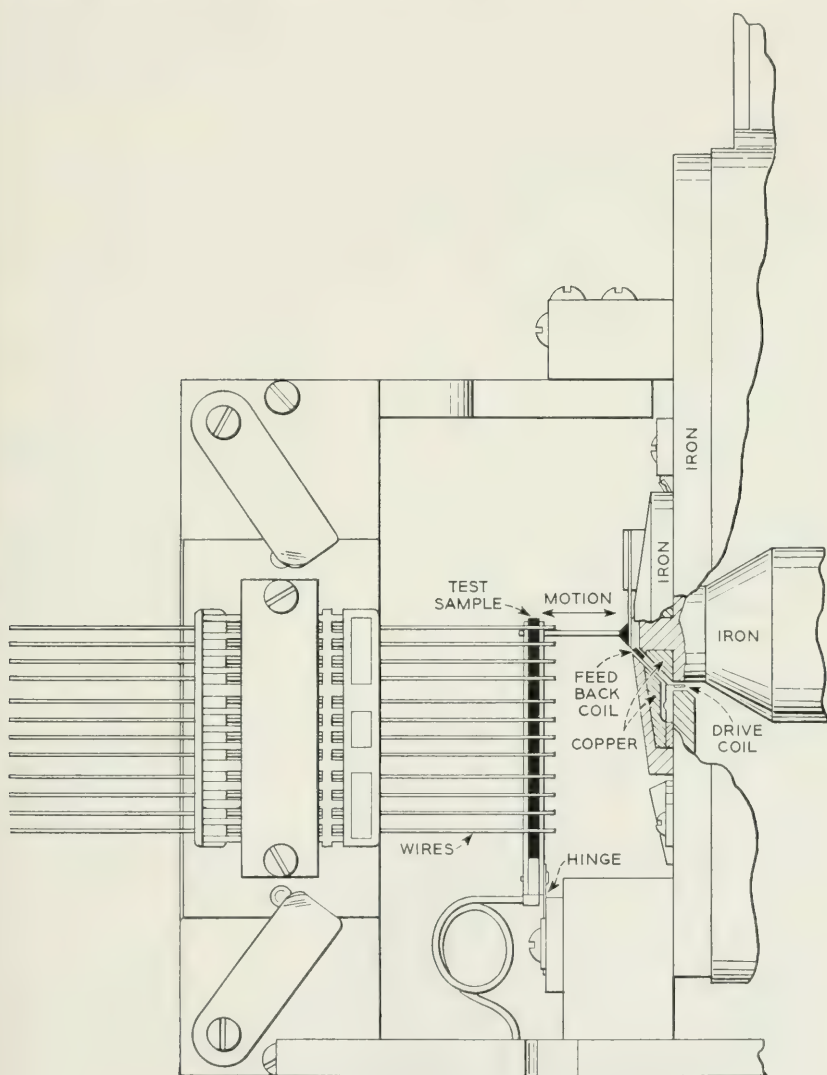


Fig. 12—Low frequency wear measuring device.

element which is carried at its base by a diaphragm and at its apex by cantilever springs. These furnish the restoring force and restrict the motion of the moving system to a single degree of freedom, motion parallel to the cone and coil axis. The second or feedback coil is secured to the cone near its apex, at which point the stylus or drive pin is attached. The coils move axially in annular air gaps polarized by a single magnet. In the space between the two coils, copper ring shielding (a shorted turn) is provided to minimize inductive coupling between them. The output of the driving amplifier is supplied to the drive coil, while the feedback coil is connected in proper (negative feedback) phase to the amplifier input.

The voltage generated in the feedback coil is proportional to the instantaneous velocity of the moving system, and by virtue of the negative feedback, the amplifier-recorder system becomes a high force, high mechanical impedance generator of mechanical motion, with the velocity very nearly proportional to the input voltage over a large range of frequency and mechanical load. Measurements of the voltage generated in the feedback coil provides a means of monitoring the velocity. Enough power capacity is present in the amplifier so that large changes in the load will not cause changes in the motion.

The samples of the materials to be tested for wear resistance are carried by a grooved aluminum beam, one end of which is hinged, the other being driven by the record stylus. The rubbing member, in this case 25-mil nickel silver wires, are tensioned against the test samples as they might be in switching apparatus. The wires can be removed for observation and measurements of the wear, and accurately replaced as the parts are dowelled together.

Fig. 13 shows a measurement of a number of materials for a normal force of 30 grams (0.0665 pounds) and a slide of 2 mil inches. The A phenolic, which is the same as that tested and recorded in Fig. 10 by the 18,000-cycle barium titanate transducer, produced essentially the same wear showing that the wear is approximately independent of the rapidity of motion for these materials. Nylon showed a rather erratic wear curve due to the fact that it has a low melting point and tends to ball up on the wires. This effect was considerably more pronounced at 18,000 cycles, where a very large indentation was found.

Only three materials show low wear at reasonably uniform rates out to a large number of cycles. These are C phenolic, a fabric filled phenolic, the B phenolic, a wood flour filled molding phenolic and the D phenolic, a cotton flock phenolic with graphite added. At lower forces and shorter slides the wear at 10^9 cycles is approximately proportional to the force

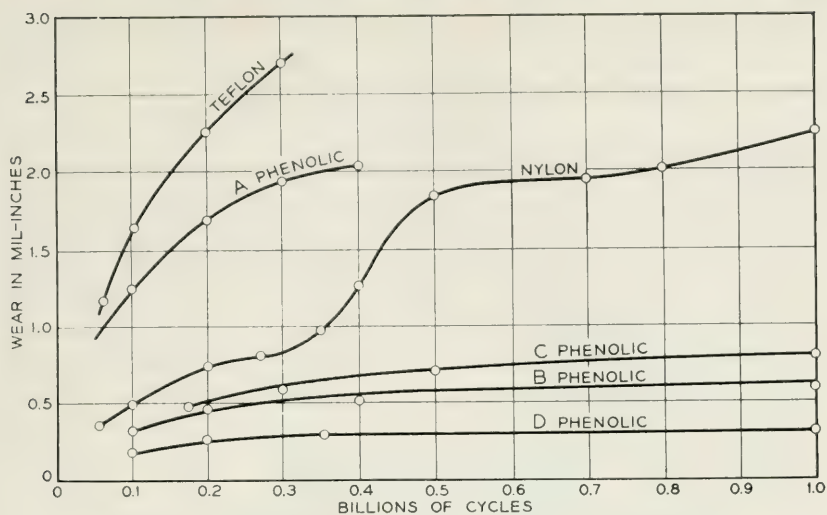


Fig. 13—Typical wear curves for a number of materials.

times the length of slide. Any of these three materials give sufficiently small wear to produce a long relay life, but the best performer under all conditions of force and slide appears to be the D cotton flock filled phenolic with graphite added.

In order to determine the causes of wear over a greater range of parameters a number of other materials were run by means of the barium titanate transducer. The wear for 2 mils motion, 30 grams (0.0665 pounds) force, and 10^9 cycles are shown by Table II.

C. Wearing Energy and Causes of Wear

A rough estimate of the energy required to break off pieces of the material shows that most of the energy goes into producing heat and very little into wear, i.e., into breaking pieces from the material. To show this let us consider a small cube fixed at one end and with a tangential force at the other. The force will cause the top surface to move with respect to the bottom surface as shown by Fig. 14, and a shearing strain S is set up in the material whose value is equal to

$$F = \mu S dx dy \quad (5)$$

where dx and dy are the cross section dimensions and μ the shear stiffness. In this displacement work is done by the sidewise displacement u equal to

$$W = \frac{1}{2} u F \quad (6)$$

But u the displacement is

$$u = \frac{\partial u}{\partial z} dz = S dz \quad (7)$$

and hence the total work done is

$$W = \frac{1}{2} \mu S^2 dx dy dz = \frac{1}{2} (\mu S^2) \times \text{volume of material} \quad (8)$$

If the force is increased, the shearing strain S increases until it reaches the limiting strain that the material can stand. This limiting strain depends on the material and whether the strain is long repeated so that the material becomes fatigued. For most plastics this limiting strain is in the order of 1 per cent and for most metals the value is less than this.

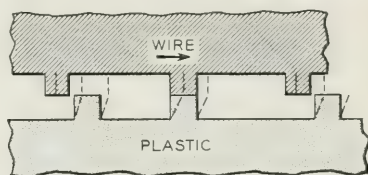


Fig. 14—Representation of points of contact and their displacements for plastic and wire.

Hence the energy to break up one cubic centimeter of material is

$$W = \frac{1}{2} \mu S_M^2 \quad (9)$$

where S_M is the breaking strain. For a plastic having a shear stiffness of $\mu = 2 \times 10^{10}$ dynes/cm² and a breaking strain of 0.01, the energy is 10^6 ergs per cubic centimeter.

This rough calculation and the amount of wear observed for various length strokes and forces allow a determination of the amount of energy going into wear production. The amount of work generated by a displacement of 0.002 inches or 0.005 cm with a normal force of 30 grams is

$$W = 0.005 \times 30 \times 980 \times f \text{ in ergs} \quad (10)$$

where f is the coefficient of friction. Since this is about 0.25 the work per stroke is 37 ergs. Twice this amount results from a complete cycle and for 10^9 cycles the work done is

$$W = 37 \times 2 \times 10^9 = 7.44 \times 10^{10} \text{ ergs} \quad (11)$$

The volume of wear observed for this condition is about 1×10^{-4} cubic cm for the A phenolic and hence we find that the part of the energy

that goes into producing wear is

$$\frac{1 \times 10^{-4} \times 10^6}{7.4 \times 10^{10}} = 1.35 \times 10^{-9} \quad (12)$$

or about 1 part in 10^9 . This suggests that the wire goes back and forth over the same high points many millions of times until the material finally becomes fatigued and breaks off. This view is confirmed by the oscillograph pictures of Fig. 9 which are a stationary pattern for millions of oscillations.

According to this picture, the material that will wear the best is the one with the highest limiting shearing strain. If we assume that the limiting shearing strain is proportional to the limiting elongation strain under repeated vibrations—of which there are tables—the wear for various materials given in Table II agrees roughly with this concept. Table II shows the yield stresses, the Young's moduli, the per cent strains at the yield point and the relative wear at 10^9 cycles. It will be seen that the materials with the highest yield strain will in general wear longer than those with smaller yield strains.

An exception to this rule was nylon which had a large wear even though it has a large yield strain. However, nylon has a relatively low softening temperature and a low heat conductivity. Observations showed that the nylon was melted off rather than abraided off. According to this rule gum rubber should wear much better than any other material since it has such a high limiting shearing strain. A run was made with a two mil inch motion on a gum rubber specimen and no observable wear was found. The fact that a rubber tire will outwear a metal tire is also confirmation of this rule.

All the tests showed that the wear on the stainless steel or nickel silver

TABLE II
AMOUNT OF WEAR FOR VARIOUS MATERIALS CAUSED BY SLIDING A 0.025 INCH
NICKEL SILVER WIRE FOR 2 MIL INCHES, 30 GRAMS
NORMAL FORCE AND 10^9 CYCLES

Material	Yield Stress Dynes/Cm ²	Youngs Modulus Dynes/Cm ²	Per Cent Yield Strain	Wear, Cubic Cm, for 2-Mil Motion for 10^9 Cycles and 30-Gram Force
Lead Glass....	$2.4 \text{ to } 2.7 \times 10^9$	6.5×10^{11}	0.0037 to 0.0041	0.027
Brass.....	$3.7 \text{ to } 4.6 \times 10^9$	9×10^{11}	0.0041 to 0.0051	0.0075
Stainless Steel.	$1.1 \text{ to } 1.4 \times 10^{10}$	2×10^{12}	0.0055 to 0.007	0.00075
B Phenolic....	7.2×10^8	6.9×10^{10}	0.0105	0.000025

wire used to produce the wear on the plastic was always very much less than that of the plastic. The reason for this as seen from Fig. 14 is that since the displacement for a given force to break the bond between two high points is going to be inversely proportional to the shearing stiffnesses of the two materials, the displacement for stainless steel with a shear stiffness of 8×10^{11} dynes/cm² will be $\frac{1}{40}$ that of the plastic with a shear modulus of 2×10^{10} dynes/cm². Hence, the shearing strain for the stainless steel is much further below its limiting strain than is the shearing strain for the plastic. When the stainless steel wire was run against a bar of synthetic sapphire—which has a much higher shear constant—the stainless steel wire was soon worn through, while little wear occurred on the sapphire.

IV. THEORETICAL AND EXPERIMENTAL INVESTIGATION OF THE NO GROSS SLIDE REGION

* Since in the no gross slide region, the shearing strain is less than in the gross slide region, the rate of wear should be considerably less. This is confirmed by direct tests of the wear as shown by Fig. 11, and by supplementary tests. Hence a further experimental and theoretical investigation has been made of this region which is defined by the condition that the tangential force is less than the product of the normal force by the coefficient of friction. If sliding motions can be kept small enough to be in this region, very little wear should occur.

Using a shear ceramic for measuring the tangential force, the static load was varied and the motion required to produce no gross slide was determined. Oscillograph figures of the type shown by Fig. 9 were used and when the figure was broadened out as shown by the third figure it was assumed that slide had occurred. Fig. 15, upper curve, shows the total motion in mil inches, plotted against the static force in grams, which will just cause gross slide. The bottom line shows the maximum shearing force in grams. This is slightly lower than the force determined by the coefficient of friction since the force becomes slightly larger as shown by the pictures of Fig. 9, when gross slide occurs. The total displacement for no slide increases as the two-thirds power of the static load.

Since neither the wire nor the plastic material is smooth, contact between the two is established at only a few points. To interpret the results obtained above, some calculations due to R. D. Mindlin¹¹ are used. These deal with the tangential forces and displacements of two balls pressed together, and are for conditions occurring before gross slide begins.

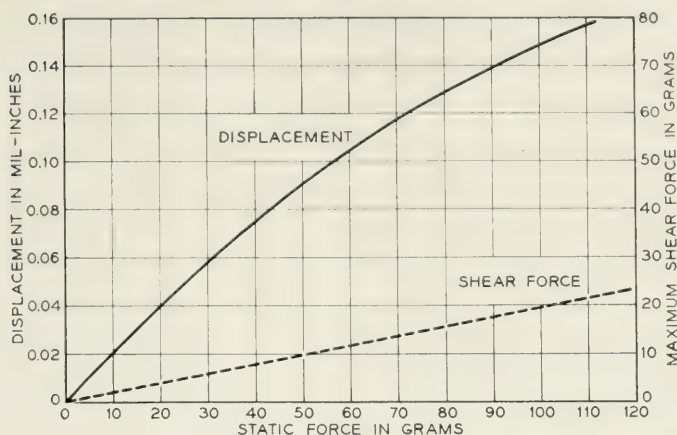


Fig. 15—Maximum total motion for no gross slide plotted against normal force. Low curve shows maximum tangential force.

From the Hertz theory of contacts,¹² the radius of contact a between two spheres is equal to

$$a = \sqrt[3]{\frac{3}{8} r N \left(\frac{1 - \sigma_1}{\mu_1} + \frac{1 - \sigma_2}{\mu_2} \right)} \quad (13)$$

where r is the radius of the spheres, N the normal force, μ_1 and σ_1 the shear elastic constant and Poisson's ratio for one sphere and μ_2 and σ_2 the same quantities for the second sphere. If now a tangential force T is applied to one of the spheres directed in the form of a couple, elastic theory shows that the tangential traction is everywhere parallel to the direction of the applied force and contours of constant tangential traction are concentric circles. The magnitude of the traction as shown by Fig. 16 rises from one half the average at the center to infinity at the edge of the circle of contact. The displacement of the circle of contact of one sphere with respect to its center is

$$\delta_x = \frac{2 - \sigma}{8\mu a} T \quad (14)$$

where a is the radius of the contact area which is given in terms of the normal force by Equation (13).

A feature of this solution that requires further study is the infinite traction at the edge of the circle of contact. Presumably the tangential component of traction cannot exceed the product of the coefficient of friction f and the normal component of traction p , which from the Hertz

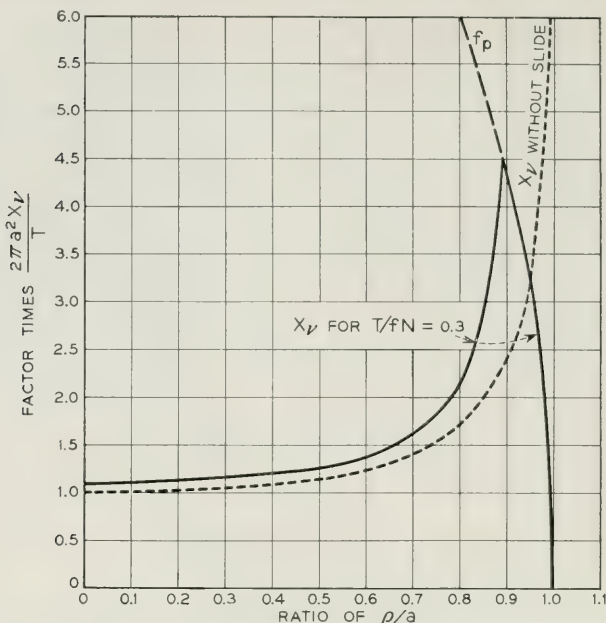


Fig. 16—Traction plotted against radius for elastic displacement and modification introduced by the effect of slip.

contact theory is

$$p = \frac{3N}{2\pi a^2} \sqrt{1 - \frac{r^2}{a^2}} \quad (15)$$

Mindlin assumes that slip takes place between the two surfaces until the tangential traction is equal to

$$X_v = \frac{3fN}{2\pi a^2} \sqrt{1 - \frac{\rho^2}{a^2}} \quad \text{for } a' \leq \rho \leq a \quad (16)$$

and less than this for all interior points, where in this equation ρ is the radius vector and a' the inner radius for which slip stops. This corresponds to the introduction of a new system of forces and Mindlin has shown that equilibrium is reestablished when the surface tractions are given by Equation (16) when $a' \leq \rho \leq a$ and by

$$X_v = \frac{3fN}{2\pi a^2} \left[\sqrt{1 - \frac{\rho^2}{a^2}} - \frac{a'}{a} \sqrt{1 - \frac{\rho^2}{a'^2}} \right] \quad \text{when } \rho \leq a' \quad (17)$$

Fig. 16 shows this distribution for the case $T/fN = 0.3$. The inner radius a' is given such a value that the integrated traction over the surface

equals T and its value is found to be

$$a' = a \sqrt[3]{1 - \frac{T}{fN}} \quad (18)$$

The added slip increases the displacement δ_x and it is shown that the total displacement is equal to

$$\delta_x = \frac{3fN(2 - \sigma)}{16\mu a} \left[1 - \left(1 - \frac{T}{fN} \right)^{2/3} \right] \quad (19)$$

A plot of this curve is shown by the line OPQ of Fig. 17 and it is evident that the displacement before gross slip occurs is 1.5 times larger than the elastic displacement calculated on the assumption of no slip.

These calculations have been extended in a recent paper³ to include the case of a cyclically varying force $T \leq fN$ and it is shown that the force displacement curve is a hysteresis type loop whose end points lie

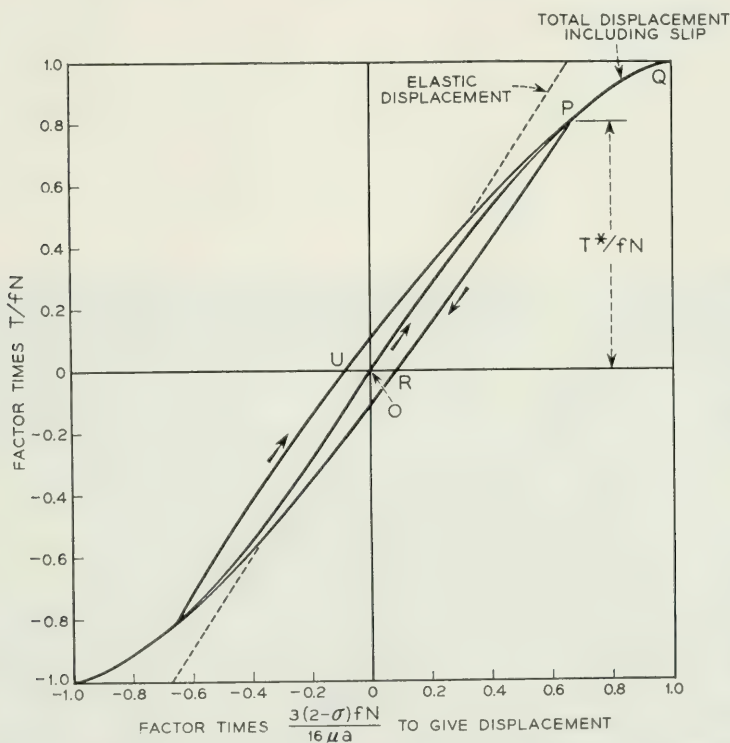


Fig. 17—Displacement versus force when slip is introduced. Hysteresis curve PRU shows displacement for an oscillating force.

on the OPQ curve of Fig. 17 and whose theoretical area W is

$$W = \frac{9(2 - \sigma) f^2 N^2}{10\mu a} \times \left[1 - \left(1 - \frac{T^*}{fN} \right)^{5/3} - \frac{5T^*}{6fN} \left[1 + \left(1 - \frac{T^*}{fN} \right)^{2/3} \right] \right] \quad (20)$$

where during the oscillation the tangential force T varies between the limits $\pm T^*$. Slip takes place as before between the radii a and a' given by

$$a' = a \sqrt[3]{1 - \frac{T^*}{fN}} \quad \text{or conversely} \quad \frac{T^*}{fN} = 1 - \frac{a'^3}{a^3} \quad (21)$$

Since the distribution of traction over the surface cannot be uniquely derived from elastic theory, the introduction of the slip function is an assumption that has to be justified by experiment. This assumption has been shown to correspond with experiment by employing the experimental arrangement shown by the photograph of Fig. 18. A barium titanate driver shown in more detail in Fig. 19 drives the middle of three glass lenses that are pressed together by a static force applied to the lever system as shown by Fig. 19. The central glass lens has a radius of curvature of 4.85 inches on each side while the other two lenses have the same radius of curvature on the sides touching the middle lens, but are flat on the other two sides and are rigidly attached to the lower platform and upper hinged lever by cement.

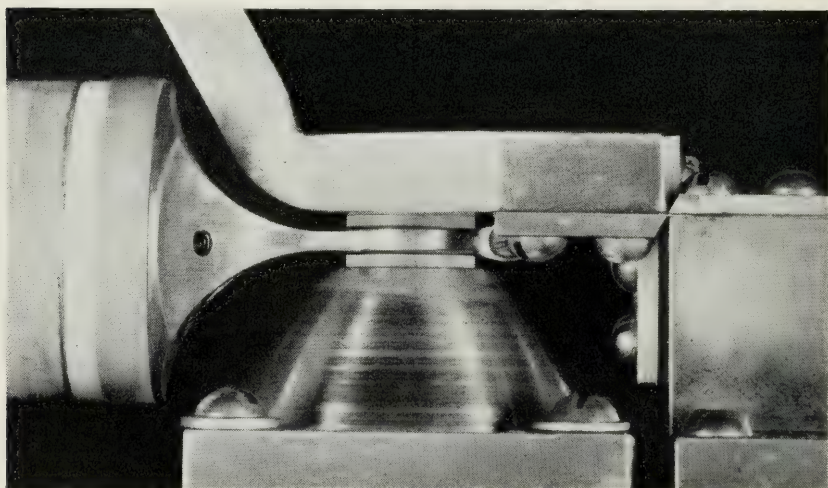


Fig. 18—Barium titanate driver, pick-up device and glass lenses.

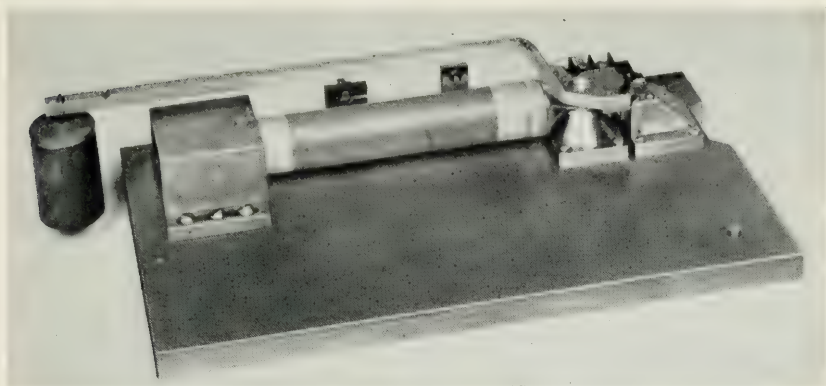


Fig. 19—The entire experimental arrangement.

circles of contact to be easily observed, normal loads on the lens system were 10 and 15 pounds which resulted in contact circles of 0.030 and 0.034 inches diameter.

With normal forces up to 15 pounds and two surfaces in contact, tangential forces up to 7.5 pounds are necessary in order to bring the central lens near the sliding point if the coefficient of friction is near one-quarter. This force was obtained by impressing voltages in the order of 3,000 rms volts on the barium titanate lead titanate hollow cylinder. This cylinder is $4\frac{3}{4}$ inches long, and has an outside diameter of 1 inch and an inside diameter of $\frac{1}{2}$ inch. The ceramic was poled in a radial direction and the constants of the material were such that a force of 167 pounds could be generated along the length for a clamped driver when a voltage of 3,000 volts (4,750 volts cm) was used. On the other hand if the driver works against no stress, the expansion in the plated length of 4 inches is 0.7×10^{-4} inches.

The actual force applied depends on how much the relative slip between the glass lenses amounts to. To measure this force, a poled lead titanate barium titanate disk is placed between the driver and the metallic bracket which clamps the middle lens as shown by Fig. 18. All the force exerted on the lens has to be exerted through the disk and hence the voltage generated by the disk is a measure of the force exerted on the middle lens. This voltage is calibrated by attaching a spring load of known constants and measuring the displacement of the load by means of a microscope.

Using a 60-cycle driving voltage, a number of sets of disks were run with varying tangential and normal loads and the wear patterns observed. Fig. 20 is a photograph (magnified 100 times) for a normal load



Fig. 20—Wear circles (magnified 100 times).

of 10 pounds and a maximum tangential load of 2.04 pounds per lens run for about 3 hours at 60 vibrations per second. The outer area of contact is seen to be 0.03 inches in diameter. The inner area of wear is a circle displaced slightly from a concentric form and has a diameter of 0.0175 inch. If we plot $1 - (a'/a)^3$ against the ratio of tangential to normal force, where a' is the inner radius and a the outer radius, as shown by Fig. 21, a point at 0.204 and 0.8 is obtained. A number of sets of lenses were run and as shown by Fig. 21 the results can be plotted on a straight line corresponding to a coefficient of friction of 0.25. This value agrees well with other determinations¹³ of the coefficient of friction of glass on glass. Hence the assumption of slip between spheres under tangential forces appears to be verified. This type of slip may be responsible for some types of wear, such as in ball bearings, where no gross slide of one surface over another occurs.

An attempt was also made to check the area of the loop as determined theoretically by Equation (20). The applied force is measured directly by the barium titanate pickup and the displacement was measured by

attaching a velocity microphone pickup to the transducer. The force voltage was placed on one set of plates of an oscillograph while the integrated output from the velocity pickup was placed on the other set. A series of oscillographs were taken for various amplitudes of motion and the pictures are shown by Fig. 22. Since the force and displacement measurements were separately calibrated, the area of the curves in inch pounds could be evaluated and are shown by Fig. 23. For amplitudes of motion near the gross slip amplitude, the area agree well with that calculated from Equation (20) from which the dotted line is obtained. For lower amplitudes the measured area is larger than the calculated area. Possibly a stick-slip process is causing the displacement to lag behind the applied force. The measured areas are nearly proportional to the square of the amplitude. The mechanical resistance associated with the stress-strain hysteresis curves of this sort is of the same type that

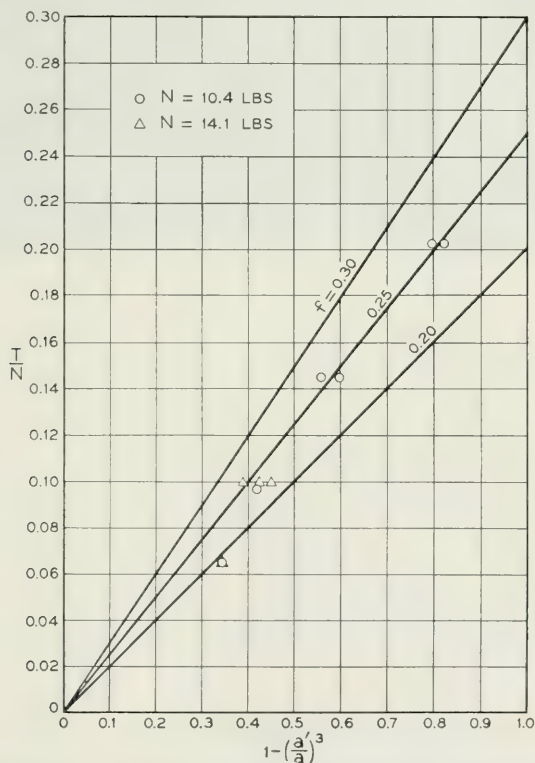


Fig. 21—Plot of $1 - (a'/a)^3$ against ratio of tangential and normal forces.

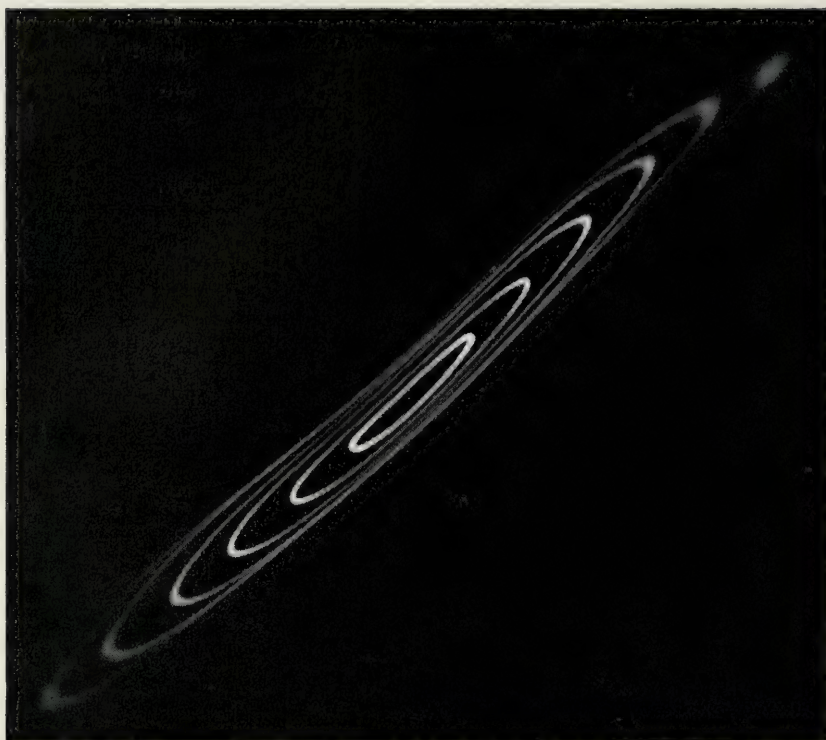


Fig. 22—Force displacement loops.

occurs in an assemblage of granular particles such as in a telephone transmitter for which the motion is so small that gross slide does not occur.

Since the theoretical displacement of Equation (19) has been verified by the glass lens experiment, we can use it to determine some of the quantities involved in the oscillographs of Fig. 9 and the displacement-normal force curve of Fig. 15. To obtain the relation between the total displacement δ_x and the normal force N , we have to eliminate a from Equation (17) since a is also a function of the normal force as shown by Equation (13). Introducing this equation, and neglecting $1/\mu_2$ as compared to $1/\mu_1$, since for the wire μ_2 is 40 times μ_1 of a plastic,

$$\delta_x = \frac{\left(\frac{3N}{\mu}\right)^{2/3} (2 - \sigma)f}{8\sqrt[3]{r(1 - \sigma)}} \left[1 - \left(1 - \frac{T^*}{fN}\right)^{2/3} \right] \quad (22)$$

Hence in agreement with the data of Fig. 15, the displacement for no gross slide should vary as the two-thirds power of the normal force.

Another deduction from Equation (22) is that the displacement for no slide should vary as the inverse two-thirds power of the shear stiffness constant μ . For example gum rubber with a shear stiffness of 2×10^7 dynes/cm² should give 100 times the displacement of a plastic with a shear stiffness of 2×10^{10} dynes/cm². A rough check of this deduction has been made by cementing a thin strip of gum rubber on the face of a shear responding ceramic and with a normal force of 30 grams (0.0665 pounds), vibrating the wire at its full amplitude of 2 mil inches. Over this range the voltage response was sinusoidal indicating that no gross slide took place. This is 33 times as large a motion as occurred for a plastic with an elastic stiffness 1,000 times that of the rubber and verifies the variation of δ_x with μ .

The other experimental quantity that can be obtained from Equation (22) is the radius r of the effective contact points of the plastic. If all

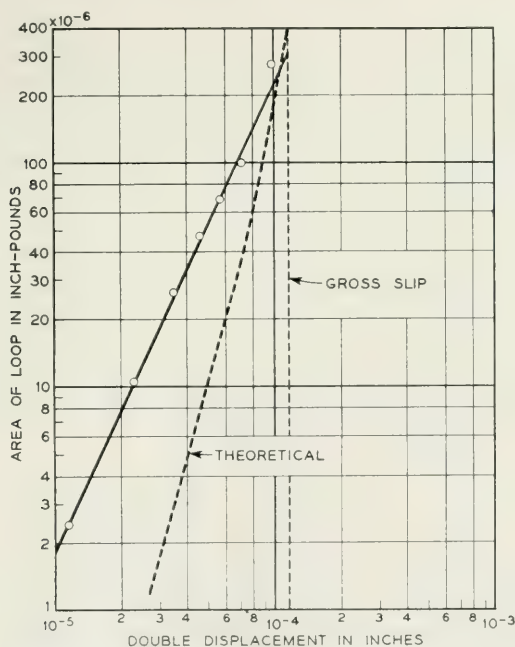


Fig. 23—Plot of area of force displacement loop against double displacement.

the force is supported by a single point at a time, then for

$$\begin{aligned}\delta_x &= \pm 3 \times 10^{-5} \text{ inches} = \pm 7.5 \times 10^{-5} \text{ cm;} \\ N &= 30 \text{ grams} = 2.94 \times 10^4 \text{ dynes;} \\ \mu &= 2 \times 10^{10} \text{ dynes/cm}^2; \\ \sigma &= 0.45\end{aligned}\tag{23}$$

and the coefficient of friction $f = 0.25$, the value of r becomes 0.008 cm. If the weight were supported equally by n points the radius would be divided by n^2 . Since the sidewise displacement would result in a strain of 0.009 for a single point and 0.036 for two points, the latter strain would be beyond the yield strain for the material. Hence the evidence seems to indicate that a single point supports the major part of the weight at any particular time.

While it is difficult to reduce the gross tangential slide of a relay to the values required for the low wear (no gross slide) region, the existence of such a region has considerable importance for other sources of wear in relays, namely long continued vibrations of component parts such as undamped wires. The tangential motions caused by such vibrations are small, but since they are repeated many times for each operation, the total integrated wear is considerable. By introducing damping so that the vibrations are quickly brought down to the low wear, no gross slide region, a considerable reduction in wear has been found for relays.

APPENDIX

VOLTAGE GENERATED BY COMPRESSIONAL AND TANGENTIAL CERAMICS BY FORCES APPLIED UNIFORMLY OR AT CONCENTRATED POINTS

When a stress is applied to a prepolarized barium titanate ceramic it has been shown¹⁴ that the open circuit field generated along the Z axis is given by the equation

$$E_3 = -2[Q_{11}[\delta_{3_0}T_3 + \delta_{1_0}T_5 + \delta_{2_0}T_4] + Q_{12}[\delta_{3_0}(T_1 + T_2) - (\delta_{1_0}T_5 + \delta_{2_0}T_4)]]\tag{24}$$

where δ_{1_0} , δ_{2_0} , δ_{3_0} are the remanent values of polarization introduced along the three axes by the poling process, T_1 , T_2 , T_3 , T_4 , T_5 , T_6 the three extensional stresses and the three shearing stresses, and Q_{11} and Q_{12} are the two electrostrictive constants for the ceramic. From the "effective" piezoelectric constants measured for these ceramics we find

that

$$Q_{11}\delta_{3_0} = 2.2 \times 10^{-8}; \quad Q_{12}\delta_{3_0} = -.8 \times 10^{-8} \text{ cgs units} \quad (25)$$

for pure barium titanate and

$$Q_{11}\delta_{3_0} = 2.4 \times 10^{-8}; \quad Q_{12}\delta_{3_0} = -.9 \times 10^{-8} \text{ cgs units} \quad (26)$$

for 4 per cent lead titanate barium titanate ceramic.

If a force F is applied uniformly over the whole surface of a small barium titanate unit, then $T_3 = F/A$, where A is the area, and all the other stresses are zero. Under these circumstances when the permanent polarization δ_{3_0} is along the Z axis (normal poling), the open circuit potential is

$$E_3 = \frac{V_3}{l_t} = \frac{2Q_{11}\delta_{3_0}F}{l_w l} = \frac{2 \times 2.4 \times 10^{-8} \times F}{l_w l} \text{ cgs units} \quad (27)$$

where l_t is the thickness and l_w and l the cross-sectional dimensions. To get the number of volts generated this factor is multiplied by 300 and

$$V_3 = \frac{1.44 \times 10^{-5} F}{l_w l} \text{ volts} \quad (28)$$

where force F is expressed in dynes.

However for the data of Figs. 2 and 4, the voltage measured is that for a load applied at the center of the ceramic and for this case the stresses T_1 and T_2 of Equation (24) cannot be neglected. The solution¹⁵ for the stresses occurring when a load F is applied at a point on the surface of a semi infinite solid is used to evaluate the corrections caused by the non-uniform load. In cylindrical coordinates the formulae for the three stresses T_{zz} and T_{rr} and $T_{\theta\theta}$ given by Timoshenko are

$$\begin{aligned} T_{rr} &= \frac{F}{2\pi} \left[(1 - 2\sigma) \left[\frac{1}{r^2} - \frac{z}{r^2} (r^2 + z^2)^{-1/2} \right] - 3r^2 z (r^2 + z^2)^{-5/2} \right] \\ T_{\theta\theta} &= \frac{F}{2\pi} (1 - 2\sigma) \left[-\frac{1}{r^2} + \frac{z}{r^2} (r^2 + z^2)^{-1/2} + z(r^2 + z^2)^{-3/2} \right] \\ T_{zz} &= \frac{-3F}{2\pi} z^3 (r^2 + z^2)^{-5/2} \end{aligned} \quad (29)$$

where r is the radial distance from the point of contact, z the distance below the surface and σ = Poisson's ratio.

The response of a barium titanate unit in terms of cylindrical coordinates has been shown¹⁶ to be for a unit polarized along the z axis

$$E_z = -2 [Q_{11}\delta_{3_0} T_{zz} + Q_{12}\delta_{3_0} (T_{rr} + T_{\theta\theta})] \quad (30)$$

Now since the ceramic is plated, the major surface is an equipotential surface and hence E_z does not vary with r or θ . Hence integrating over the surface of the ceramic, we have for the open circuit field

$$E_z \int_0^\infty \int_0^{2\pi} r \, dr \, d\theta = -2 \left[Q_{11}\delta_{30} \int_0^\infty \int_0^{2\pi} T_{zz}r \, dr \, d\theta + Q_{12}\delta_{30} \int_0^\infty \int_0^{2\pi} (T_{rr} + T_{\theta\theta})r \, dr \, d\theta \right] \quad (31)$$

Introducing the values of T_{zz} , T_{rr} and $T_{\theta\theta}$ from Equation (29) and performing the integrations we find

$$E_z A = 2 [Q_{11}\delta_{30}F + Q_{12}\delta_{30} (1 + 2\sigma)F] \quad (32)$$

where A is the cross-sectional area of the ceramic. The first term agrees with that for a uniform stress, but the second term shows that we have a correction due to the radial and tangential stresses generated by the application of the force at a point.

The amount of correction can be calculated by putting in the values of Q_{12} and σ the Poisson ratio. Recent measurements of the thickness resonance and the resonance of a torsional ceramic have shown that the best values of the Lamé elastic constants are

$$\lambda = 5.8 \times 10^{11} \text{ dynes/cm}^2; \quad \mu = 4 \times 10^{11} \text{ dynes/cm}^2 \quad (33)$$

With these values, Poisson's ratio becomes

$$\sigma = \frac{\lambda}{2(\lambda + \mu)} = \frac{5.8}{19.6} = 0.296 \quad (34)$$

For 4 per cent lead titanate barium titanate ceramic, introducing the values given above, the voltage generated by a force applied at a point is about 0.4 of that for a force applied uniformly, giving

$$V_z = \frac{0.575 \times 10^{-5} F l_t}{l_w l} \text{ volts} \quad (35)$$

This value corresponds reasonably well with the data of Fig. 4.

When the remanent polarization is applied along the Y axis and the voltage measured along the Z axis, Equation (24) shows that the open circuit voltage will be

$$E_3 = -2 (Q_{11} - Q_{12})\delta_{20}T_4 \quad (36)$$

where $T_4 = Y_z$ is the stress in the direction of polarization (Y) applied to the surface of the ceramic. Since the single stress T_4 is involved, the

open circuit voltage will be independent of whether the force is applied uniformly over the surface or at a point. This follows from the fact that E_3 is independent of x and y and hence

$$E_3 \int_0^{lw} \int_0^l dx dy = -2(Q_{11} - Q_{12})\delta_{20} \int_0^{lw} \int_0^l T_4 dx dy \quad (37)$$

Integrating over the surface gives the total force F for the right side and hence

$$\begin{aligned} V_3 &= \frac{2(Q_{11} - Q_{12})\delta_{20}l_t F}{l_w l} \text{ in cgs units} \\ &= \frac{1.98 \times 10^{-5} l_t F}{l_w l} \text{ in volts} \end{aligned} \quad (38)$$

For a ceramic 0.1 cm by 0.1 cm in cross-section and 0.05 cm thick a tangential force of 100 grams should generate a voltage of 9.7 volts.

REFERENCES

1. L. Vieth and C. F. Wiebusch, "Recent Developments in Hill and Dale Recorders," *J. Soc. Motion Pictures Engrs.*, Jan., 1938.
2. W. P. Mason and R. F. Wick, "A Barium Titanate Transducer Capable of Large Motion at an Ultrasonic Frequency," *J. Acous. Soc. of A.*, **23**, pp. 209-214, Mar., 1951.
3. R. D. Mindlin, W. P. Mason, T. F. Osmer, and H. Deresiewicz, "Effects of an Oscillating Tangential Force on the Contact Surfaces of Elastic Spheres," presented before First National Congress of Applied Mechanics, June 14, 1951. The results of this paper are summarized here.
4. Measurements have been made by T. F. Osmer.
5. This circuit was devised by G. A. Head.
6. W. P. Mason, *Piezoelectric Crystals and Their Application to Ultrasonics*, D. Van Nostrand, Chapter XII, 1950.
7. The data of Figs. 2 and 4 were obtained by L. Egerton.
8. Elizabeth A. Wood, "Detwinning Ferroelectric Crystals," *Bell System Tech. J.*, **30**, No. 4, Part I, pp. 945-955, Oct., 1951.
9. W. P. Mason, *Piezoelectric Crystals and Their Application to Ultrasonics*, D. Van Nostrand, Chap. XII, 1950.
10. The photographs of Fig. 6 were obtained by T. E. Davis.
11. R. D. Mindlin, "Compliance of Elastic Bodies in Contact," *J. Appl. Mech.*, pp. 259-268, September, 1949.
12. A. E. H. Love, *Theory of Elasticity*, 4th Edition, page 198, Cambridge University Press.
13. I. Simon, O. McMahon and R. J. Bowen, "Dry Metallic Friction as a Function of Temperature Between 4.2°K and 600°K.," *J. App. Phys.*, **22**, pp. 170-184, Feb., 1951.
14. W. P. Mason, *Piezoelectric Crystals and Their Application to Ultrasonics*, D. Van Nostrand, Chap. XII, p. 300.
15. S. Timoshenko, *Theory of Elasticity*, McGraw-Hill Co., p. 311.
16. W. P. Mason, *Piezoelectric Crystals and Their Application to Ultrasonics*, D. Van Nostrand, Appendix A9, p. 490.

A Comparison of Signalling Alphabets

By E. N. GILBERT

(Manuscript received March 24, 1952)

Two channels are considered; a discrete channel which can transmit sequences of binary digits, and a continuous channel which can transmit band limited signals. The performance of a large number of simple signalling alphabets is computed and it is concluded that one cannot signal at rates near the channel capacity without using very complicated alphabets.

INTRODUCTION

C. E. Shannon's encoding theorems¹ associate with the channel of a communications system a capacity C . These theorems show that the output of a message source can be encoded for transmission over the channel in such a way that the rate at which errors are made at the receiving end of the system is arbitrarily small provided only that the message source produces information at a rate less than C bits per second. C is the largest rate with this property.

Although these theorems cover a wide class of channels there are two channels which can serve as models for most of the channels one meets in practice. These are:

1. *The binary channel*

This channel can transmit only sequences of binary digits 0 and 1 (which might represent hole and no hole in a punched tape; open-line and closed line; pulse and no pulse; etc.) at some definite rate, say one digit per second. There is a probability p (because of noise, or occasional equipment failure) that a transmitted 0 is received as 1 or that a transmitted 1 is received as 0. The noise is supposed to affect different digits independently. The capacity of this channel is

$$C = 1 + p \log p + (1 - p) \log (1 - p) \quad (1)$$

bits per digit. The log appearing in Equation (1) is log to the base 2; this convention will be used throughout the rest of this paper.

¹ C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Tech. J.*, **27**, p. 379-423 and pp. 623-656, 1948, theorems 9, 11, and 16 in particular.

2. The low-pass filter

The second channel is an ideal low-pass filter which attenuates completely all frequencies above a cutoff frequency W cycles per second and which passes frequencies below W without attenuation. The channel is supposed capable of handling only signals with average power P or less. Before the signal emerges from the channel, the channel adds to it a noise signal with average power N . The noise is supposed to be white Gaussian noise limited to the frequency band $|\nu| < W$. The capacity of this channel is

$$C = W \log \left(1 + \frac{P}{N} \right) \quad (2)$$

bits per second.

Shannon's theorems prove that encoding schemes exist for signalling at rates near C with arbitrarily small rates of errors without actually giving a constructive method for performing the encoding. It is of some interest to compare encoding systems which can easily be devised with these ideal systems. In Part I of this paper some schemes for signalling over the binary channel will be compared with ideal systems. In Part II the same will be done for the low-pass filter channel.

PART I

THE BINARY CHANNEL

1. Error-Correcting Alphabets

Imagine the message source to produce messages which are sequences of letters drawn from an alphabet containing K letters. We suppose that the letters are equally likely and that the letters which the source produces at different times are independent of one another. (If the source given is a finite state source which does not fit this simple description, it can be converted into one which approximately does by a preliminary encoding of the type described in Shannon's Theorem 9.) To transmit the message over the binary channel we construct a new alphabet of K letters in which the letters are different sequences of binary digits of some fixed length, say D digits. Then the new alphabet is used as an encoding of the old one suitable for transmission over the channel. For example, if the source produced sequences of letters from an alphabet of 3 letters, a typical encoding with $D = 5$ might convert the message

into a binary sequence composed of repetitions of the three letters.

00000
11100
and 00111

If $K = 2^D$, the alphabet consists of all binary sequences of length D and hence if any of the digits of a letter is altered by noise the letter will be misinterpreted at the receiving end of the channel. If K is somewhat smaller than 2^D it is possible to choose the letters so that certain kinds of errors introduced by the noise do not cause a misinterpretation at the receiver. For example, in the three letter alphabet given above, if only one of the five digits is incorrect there will be just one letter (the correct one) which agrees with the received sequence in all but one place. More generally if the letters of the alphabet are selected so that each letter differs from every other in at least $2k + 1$ out of the D places, then when k or fewer errors are made the correct interpretation of the received sequence will be the (unique) letter of the alphabet which differs from the received sequence in no more than k places. An alphabet with this property will be called a *k error correcting alphabet*².

Error correcting alphabets have the advantage over the random alphabets which Shannon used to prove his encoding theorems that they are uniformly reliable whereas Shannon's alphabets are reliable only in an average sense. That is, Shannon proved that the probability that a letter *chosen at random* shall be received incorrectly can be made arbitrarily small. However, a certain small fraction of the letters of Shannon's alphabets are allowed a much higher probability of error than the average. This kind of alphabet would be undesirable in applications such as the signalling of telephone numbers; one would not want to give a few subscribers telephone numbers which are received incorrectly more often than most of the others. It is only conjectured that the rate C can be approached using error correcting alphabets. The alphabets which are to be considered here are all error correcting alphabets.

A geometric picture of an alphabet is obtained by regarding the D digits of a sequence as coordinates of a point in Euclidean D dimensional space. The possible received sequences are represented by vertices of the unit cube. A k error correcting alphabet is represented by a set of vertices, such that each pair of vertices is separated by a distance at least $\sqrt{2k + 1}$.

Let $K_0(D, k)$ be the largest number of letters which a D dimensional

² R. W. Hamming, "Error Detecting and Error Correcting Codes," *Bell System Tech. J.*, **29**, pp. 147-160, 1950.

k error correcting alphabet can contain. Except when $k = 1$, there is no general method for constructing an alphabet with $K_0(D, k)$ letters, nor is $K_0(D, k)$ known as a function of D and k . Crude upper and lower bounds for $K_0(D, k)$ are given by the following theorem.

Theorem 1. The largest number of letters $K_0(D, k)$ satisfies

$$\frac{2^D}{N(D, 2k)} \leq K_0(D, k) \leq \frac{2^D}{N(D, k)} \quad (3)$$

where

$$N(D, k) = \sum_{r=0}^k C_{D, r}$$

is the number of sequences of D digits which differ from a given sequence in $0, 1, \dots$, or k places.

Proof

The upper bound is due to R. W. Hamming and is proved by noting that for each letter S of a k error correcting alphabet there are $N(D, k)$ possible received sequences which will be interpreted as meaning S . Hence $N(D, k) K_0(D, k) \leq 2^D$, the total number of sequences.

The lower bound is proved by a random construction method. Pick any sequence S_1 for the first letter. There remain $2^D - N(D, 2k)$ sequences which differ from S_1 in $2k + 1$ or more places. Pick any one of these S_2 for the second letter. There remain at least $2^D - 2N(D, 2k)$ sequences which differ from both S_1 and S_2 in $2k + 1$ or more places. As the process is continued, there remain at least $2^D - rN(D, 2k)$ sequences, which differ in $2k + 1$ or more places from S_1, \dots, S_r , from which S_{r+1} is chosen. If there are no choices available after choosing S_K , then $2^D - KN(D, 2k) \leq 0$ so the alphabet (S_1, \dots, S_K) has at least as many letters as the lower bound (3).

For all the simple cases (D and k not very large) investigated so far the upper bound is a better estimate of $K_0(D, k)$ than the lower bound. The upper and lower bounds differ greatly, as may be seen from a quick inspection of Table I. For example, in the case of a ten dimensional two error correcting alphabet, the bounds are 2.7 and 18.3.

2. Efficiency Graph

The first step in constructing an efficiency graph for comparing alphabets is to decide on what constitutes reliable transmission. The criterion used here is that on the average no more than one letter in 10^4 shall be misinterpreted.

TABLE I
TABLE OF $2^D/N(D, k)$

$k =$	1	2	3	4	5	6	7
$D = 3$	2						
4	3.2						
5	5.3	2					
6	9.1	2.9					
7	16	4.4	2.9				
8	28.4	6.9	2.8				
9	51.2	11.1	3.9	2			
10	93.1	18.3	5.8	2.7			
11	170.7	30.6	8.8	3.6	2		
12	315.8	51.8	13.7	5.2	2.6		
13	585.2	89.0	21.6	7.5	3.4	2	
14	1092.3	154.4	34.9	11.1	4.7	2.5	
15	2048	270.8	56.8	16.8	6.6	3.3	2

Missing entries are numbers between 1 and 2.

This sort of criterion might be appropriate for a channel transmitting English text. For other messages it is not always appropriate. For example, if the messages are telephone numbers, one would naturally require that the probability of mistaking a telephone number be small, say less than 10^{-4} . If the telephone numbers are L decimal digits long, and if the alphabet has K different letters in it (so that it takes about $L \log 10 / \log K$ letters to make up a telephone number) the probability of making a mistake in a single letter should be required to be less than about

$$\frac{10^{-4} \log K}{L \log 10}$$

which gives alphabets with large K an advantage over alphabets with small K .

Since the probability that exactly r binary digits out of D shall be received incorrectly is $C_{D,r} p^r (1-p)^{D-r}$, we achieve the required reliability with a D -dimensional k -error correcting alphabet provided p satisfies

$$\sum_{r=k+1}^D C_{D,r} p^r (1-p)^{D-r} \leq 10^{-4}. \quad (4)$$

The value of p which makes the inequality hold with the equals sign determines the noisiest channel over which the alphabet can be used safely.

Let K be the number of different letters in the alphabet. Then the

rate in bits per digit at which information is being received is

$$R = \frac{\log K}{D}. \quad (5)$$

In Equation (5) we have neglected a term which takes account of the information lost due to channel noise. This is legitimate because all but 10^{-4} of the letters are received correctly.

The worst tolerable probability p of (4) and the rate R of Equation (5) determine the noise combating ability of an alphabet. To compare different alphabets one may represent them as points on an efficiency graph of R versus p . Fig. 1 is an efficiency graph on which the values (p, R) for a number of simple error correcting alphabets have been plotted. Each point on the graph is labelled with the two numbers k, D in that order. The alphabets represented were not found by any systematic process and are not all proved to be best possible (i.e., to have the largest K) for the stated values of k and D . Fortunately, R depends on K only logarithmically so that it is not likely the points representing the best possible alphabets lie far away from the plotted points.

The solid line represents the curve

$$R = C = 1 + p \log p + (1 - p) \log (1 - p).$$

According to Shannon's theorems, all alphabets are represented by points lying below this line.

The efficiency graph only partially orders the alphabets according to

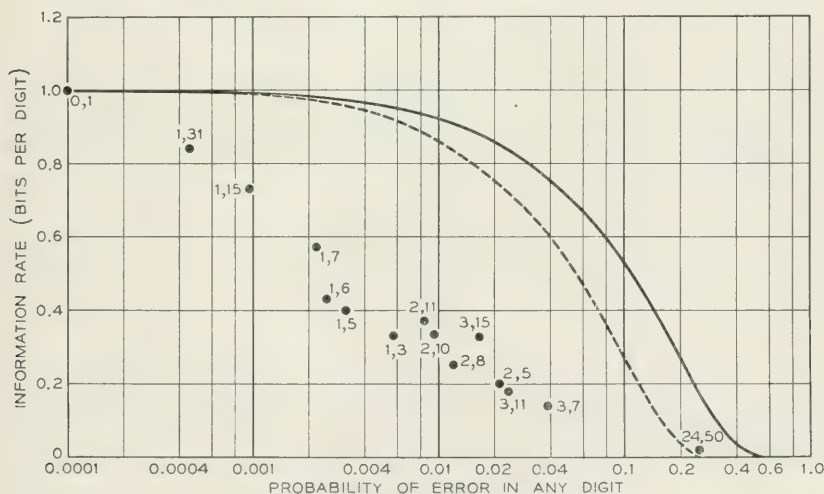


Fig. 1—Probability of error in a letter is 10^{-4} .

their invulnerability to noise. For example, it is clear that the alphabet 3, 15 is better than 2, 8. However, without further information about the channel, such as knowledge of p , there is no reasonable way of choosing between 3, 15 and 3, 7.

3. Large Alphabets

We have been unable to prove that there are error correcting alphabets which signal at rates arbitrarily close to C while maintaining an arbitrarily small probability of error for any letter. A result in this direction is the following theorem.

Theorem 2. Let any positive ϵ and δ be given. Given a channel with $p < \frac{1}{4}$ there exists an error correcting alphabet which can signal over the channel at a rate exceeding $R_0 - \epsilon$ where

$$R_0 = 1 + 2p \log 2p + (1 - 2p) \log (1 - 2p)$$

bits per digit and for which the probability of error in any letter is less than δ .

Proof

The probability of error in any letter is the sum on the left of (4). This is a sum of terms from a binomial distribution which, as is well known, tends to a Gaussian distribution with mean Dp and variance $Dp(1 - p)$ for large D . Hence there is a constant $A(\delta)$ such that all k error correcting alphabets with sufficiently large D have a letter error probability less than δ provided

$$k \geq Dp + A(\delta) (Dp(1 - p))^{1/2} \quad (6)$$

Let $k(D)$ be the smallest integer which satisfies (6) and consider an alphabet which corrects $k(D)$ errors and contains $K_0(D, k(D))$ letters. By Equation (5) and the lower bound of Theorem 1, this alphabet signals at a rate $R(D)$ satisfying

$$1 - \frac{1}{D} \log N(D, 2k(D)) \leq R(D).$$

Since $p < \frac{1}{4}$, $2k(D) < D/2$ for large D and hence

$$N(D, 2k(D)) < (2k(D) + 1)C_{D, 2k(D)}.$$

Then an application of Stirling's approximation for factorials shows that as $D \rightarrow \infty$

$$1 - \frac{1}{D} \log N(D, 2k(D)) \rightarrow R_0.$$

Hence by taking D large enough one obtains an alphabet with rate exceeding $R_0 - \epsilon$ and letter error probability less than δ .

The rate R_0 appears on the efficiency graph as a dotted line.

It has not been shown that no error-correcting alphabet has a rate exceeding R_0 . In fact, one alphabet which exceeds R_0 in rate is easy to construct. If the noise probability p is greater than $\frac{1}{4}$, then $R_0 = 0$. The alphabet with just two letters

$$0\ 0\ 0\ 0\ \dots\ 0$$

and

$$1\ 1\ 1\ 1\ \dots\ 1$$

will certainly transmit information at a (small) positive rate, and with a 10^{-4} probability of errors if D is large enough, as long as $p < \frac{1}{2}$.

Using a more refined lower bound for $K_0(D, k)$ it might be shown that there are error-correcting alphabets which signal with rates near C . If one repeats the calculation that led to R_0 using the upper bound (3) (which seems to be a better estimate of the true $K_0(D, k)$) instead of the lower bound (3), one is led to the rate C instead of R_0 .

The condition (4) is more conservative than necessary. The structure of the alphabet may be such that a particular sequence of more than k errors may occur without causing any error in the final letter. This is illustrated by the following simple example due to Shannon: the alphabet with just two letters

$$\begin{array}{cccccc} 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \end{array}$$

corrects any single error but also corrects certain more serious errors such as receiving 0 0 1 1 1 1 for 0 0 0 0 0 0. An alphabet designed for practical use would make efficient enough use of the available sequences so that any sequence of much more than k errors causes an error in the final letter; the random alphabets constructed above probably do not. If this kind of error were properly accounted for, the rate R_0 could be improved, perhaps to C .

4. Other Discrete Channels

If instead of transmitting just 0's and 1's the channel can carry more digits

$$0, 1, 2, \dots, n$$

a similar theory can be worked out. The simplest kind of noise in this channel changes a digit into any one of the n other possible numbers with probability p/n . Then the capacity of the channel is

$$C = \log(n + 1) + p \log \frac{p}{n} + (1 - p) \log(1 - p).$$

Error-correcting alphabets for this channel can also be constructed and the criterion (4) for good transmission remains unchanged. The proof of theorem 1 can be repeated with little change using

$$N(D, k) = \sum_{r=0}^k C_{D, r} n^r$$

as the number of sequences which can be reached after k or fewer errors [the terms 2^D in (1) and (3) are replaced by $(n + 1)^D$]. Once more, using the lower bound, one finds an expression for R_0 which is the same as the one for C but with p replaced by $2p$.

PART II

THE LOW PASS FILTER

1. Encoding and Detection

If $f(t)$ is a signal emerging from a low pass filter (so that its spectrum is confined to the frequency band $|\nu| < W$ cycles per second) then $f(t)$ has a special analytic form given by the sampling theorem³

$$f(t) = \sum_{m=-\infty}^{\infty} f\left(\frac{m}{2W}\right) \frac{\sin \pi(2Wt - m)}{\pi(2Wt - m)} \quad (7)$$

Thus the signal is completely determined by the sequence of sample values $f(m/2W)$. The average power of the signal $f(t)$ is measured by

$$P = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T f^2(t) dt$$

which can be expressed in terms of the sample values as follows

$$P = \lim_{M \rightarrow \infty} \frac{1}{2M} \sum_{m=-M}^M f^2\left(\frac{m}{2W}\right). \quad (8)$$

As in Part I, consider a message source producing a sequence of letters from an alphabet of K equally likely letters. To transmit this information over the low pass filter we must encode the sequence into a function

³ C. E. Shannon, "Communication in the Presence of Noise," *Proc. I. R. E.*, **37**, pp. 10-21, Jan. 1949.

$f(t)$ of the form (7), or in other words into a sequence of sample values $f(m/2W)$. To do this, we construct a new alphabet containing K letters which are different sequences of real numbers of some fixed length, say D places. When we let the letters of the new alphabet correspond to letters of the old one the message is translated into a sequence of real numbers which we use for the sequence $f(m/2W)$.

If the K letters of the sequence alphabet are

$$\begin{aligned} S_1: & a_{11}, \dots, a_{1D} \\ S_2: & a_{21}, \dots, a_{2D} \\ & \cdot \quad \cdot \quad \cdot \\ & \cdot \quad \cdot \quad \cdot \\ & \cdot \quad \cdot \quad \cdot \\ S_K: & a_{K1}, \dots, a_{KD}, \end{aligned}$$

the expression (8) for the average power of the function $f(t)$ becomes

$$P = \frac{1}{DK} (d_1^2 + d_2^2 + \dots + d_K^2) \quad (9)$$

where

$$d_i^2 = \sum_{j=1}^D a_{ij}^2.$$

If the D numbers in the sequence S_i are regarded as coordinates of a point in Euclidean D dimensional space, d_i^2 represents the square of the distance from the point representing S_i to the origin.

When $f(t)$ is transmitted, the received signal will be $f(t) + n(t)$ where $n(t)$ is some (unknown) white Gaussian noise signal. The noise signals $n(t)$ are characterized by the fact that their sample values $n(m/2W)$ are independently distributed according to Gaussian laws. That is,

$$\text{Prob} \left(n \left(\frac{m}{2W} \right) \leq X \right) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^X e^{-y^2/2\sigma^2} dy. \quad (10)$$

The variance σ^2 of the distribution of noise samples is, by an application of (8), the power of this ensemble of noise signals.

At the receiving end of the channel, there is a detector which observes each block of D sample values $f(m/2W) + n(m/2W)$ and tries to decide which one of the K letters S_1, \dots, S_K was sent. In terms of the geometric picture, the detector divides all of D dimensional space into K non-overlapping regions U_1, \dots, U_K with the property that, if the D received sample values are represented by a point in U_i , the detector

decides that S_i was sent. By Equation (10), the probability that the detector picks the wrong letter when S_i is sent is

$$p_i = \frac{1}{(2\pi)^{D/2} \sigma^D} \int \int_{\bar{U}_i} \cdots \int e^{-r_i^2/2\sigma^2} dy_1 \cdots dy_D \quad (11)$$

where \bar{U}_i is the set of all points not in U_i and r_i is the distance from (y_1, \cdots, y_D) to the point representing S_i .

For any given alphabet the best possible detector (in the sense that it minimizes the average probability of making an error in guessing a letter) is called a *maximum likelihood detector*. The region U_i for a maximum likelihood detector consists of all points (y_1, \cdots, y_D) which are closer to the point S_i than to any other letter point S_j ($r_i < r_j$ for all $j \neq i$). To prove that this choice of U_i is best possible consider any other detector such that U_i contains a set V of points in which $r_i > r_j$. A direct calculation shows that the detector obtained by removing V from U_i and making V part of U_j has a smaller probability of error per letter. The set of points equidistant from two given points is a hyperplane. The region U_i of a maximum likelihood detector is a convex region bounded by segments of the hyperplanes

$$r_i = r_1, \quad r_i = r_2, \cdots.$$

To compare signalling alphabets under the most favorable possible circumstances, we always compute letter error probabilities assuming that the detector is a maximum likelihood detector.

2. Computation of error probabilities

Exact evaluation of the letter error probability integral (11) is impossible except in a few special cases. Fortunately we are only interested in (11) when σ is small enough in comparison to the size of U_i to make the integral small. Then fairly accurate approximate formulas can be derived.

Theorem 3. Let R_{ij} be the distance between letter points S_i and S_j . Then

$$1 - \prod_{j \neq i} (1 - Q_{ij}) \leq p_i \leq \sum_{j \neq i} Q_{ij} \quad (12)$$

where

$$Q_{ij} = \frac{1}{\sqrt{2\pi}} \int_{R_{ij}/2\sigma}^{\infty} e^{-x^2/2} dx.$$

The proof of Theorem 3 follows from the fact that Q_{ij} is the probability that, when S_i is transmitted, the received sequence will be closer to S_j than to S_i .

In the cases to be computed Q_{ij} is a rapidly decreasing function of R_{ij} and the only terms worth keeping in (12) are the ones for which R_{ij} is the smallest of the numbers R_{i1}, \dots, R_{iK} . Moreover since the Q_{ij} are all small enough so that the upper and lower bounds differ only by a few per cent, the upper bound is a good approximation to p_i . Then a simple approximate formula for the average letter error probability $p = (p_1 + \dots + p_K)/K$ is

$$p = \frac{N}{\sqrt{2\pi}} \int_{r_0/\sigma}^{\infty} e^{-x^2/2} dx \quad (13)$$

where $2r_0$ is the smallest of the $K(K-1)/2$ distances R_{ij} and N is the average over all letters in the alphabet of the number of letter points which are a distance $2r_0$ away.

3. Efficiency graph

The efficiency graph to be described was constructed originally to compare alphabets for signalling telephone numbers of length equal to ten decimal digits. It was desired that on the average only one telephone number in 10^4 should be received incorrectly. As described in Part I section 2, if the telephone numbers are encoded into sequences of letters from an alphabet of K letters, we must require that the average probability of error in any letter be

$$p = 10^{-5} \log_{10} K \quad (14)$$

or smaller.

Given an alphabet, one can compute with the help of (13) and (14) and a table of the error integral the largest value of the noise power σ^2 which can be tolerated. The average power of the transmitted signal is P given by Equation (9). Hence we can compute the smallest signal to noise ratio

$$Y = P/\sigma^2 \quad (15)$$

which will be satisfactory.

A letter containing $\log K$ bits of information is transmitted during an interval of $D/2W$ seconds. Hence the rate at which information is received is

$$R = \frac{2W \log K}{D} \quad (16)$$

bits per second. Again Equation (16) ignores a term representing in-

formation lost due to channel noise which is negligible because the error probability is low.

The efficiency graph, Fig. 2, is a chart on which the signal to noise ratio Y in db [computed from Equation (15)] is plotted against the signalling rate per unit bandwidth $R/W = (2 \log K)/D$ for different alpha-

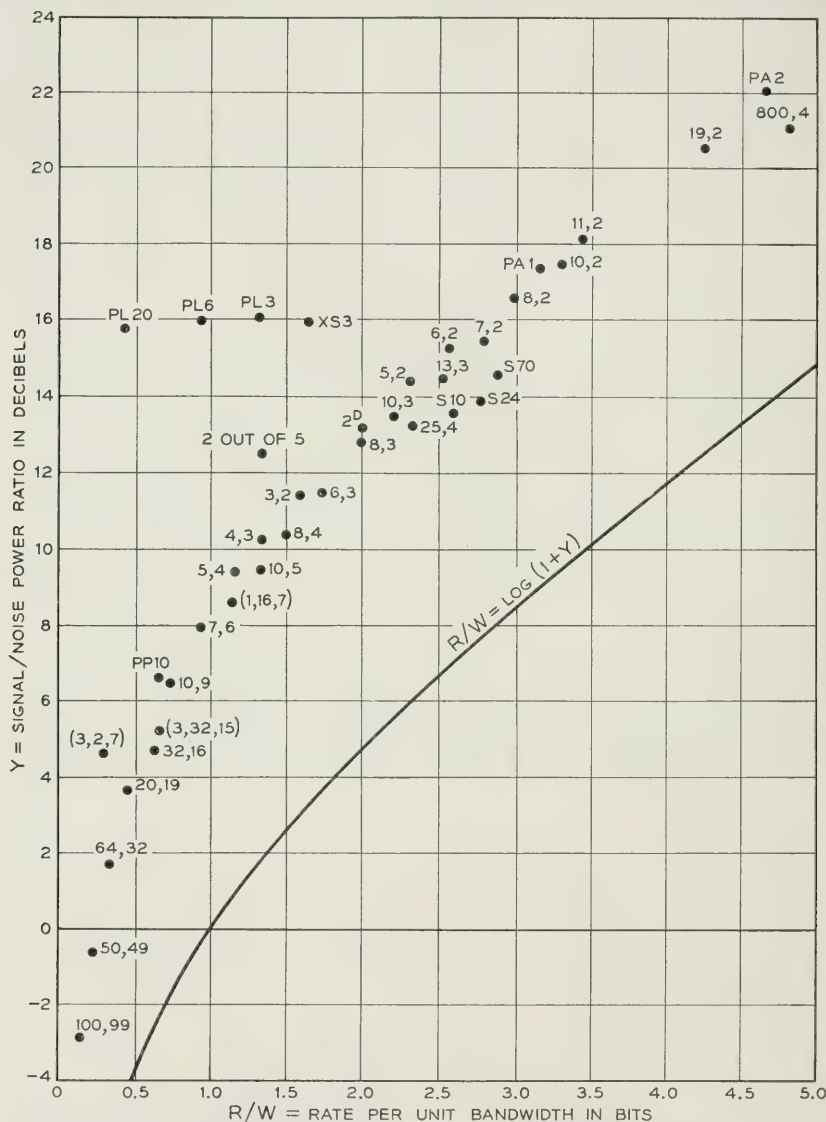


Fig. 2—Probability is 10^{-4} that an error is made in a 10 digit decimal number.

bets. An alphabet is considered poor if its point on the efficiency graph lies far above the ideal curve $R/W = C$, $W = \log(1 + Y)$.

4. The alphabets

The alphabets which appear on the efficiency graph are the following:

excess three (XS3): the ten sequences of 4 binary digits which represent 3, 4, \dots , and 12 in binary notation;

two out of five: the ten sequences of five binary digits which contain exactly two ones;

pulse position (PP10): the ten sequences of ten binary digits which contain exactly one one;

2^D *binary*: all of 2^D sequences of D binary digits.

pulse amplitude (PAn): the $2n + 1$ sequences of length 1 consisting of $-n, -n + 1, \dots, n$. This alphabet gives rise to a sort of quantized amplitude modulation.

pulse length (PLn): the $n + 1$ sequences of n binary digits of the form $11 \dots 10 \dots 0$, i.e., a run of ones followed by a run of zeros.

Minimizing alphabets (K, D): The above alphabets are taken from actual practice. They are convenient because, aside from PAn, they require a signal generator with only two amplitude levels. If we ignore ease of generating the signals as a factor, a great many geometric arrangements of points suggest themselves as possible good alphabets. The principle by which one arrives at good alphabets may be described as follows. When a D and K have been determined which give the desired information rate R [by Equation (16)] try to arrange the K letter points in D dimensional space in such a way that the distances between pairs of points are all greater than some fixed distance and that the average of the K squared distances to the origin is minimized. By Equations (9) and (13) it is seen that, apart from the small influence of the factor N , this process must minimize the signal to noise ratio Y required.

Ordinarily it is difficult to prove that a configuration is a minimizing one. Even to recognize a configuration which leads to a relative minimum (i.e. a minimum over all nearby configurations) is not always easy. The eight vertices of a cube, for example, do not give a relative minimum. Consequently, most of the alphabets to be described are only conjectured to be "best possible." Each of them satisfies one necessary requirement of minimizing alphabets that the centroid of the point configuration (assuming a unit mass at each letter point) lies at the origin. That this condition is necessary follows from the easily derived identity

$$r_2^2 = r_1^2 - R_0^2$$

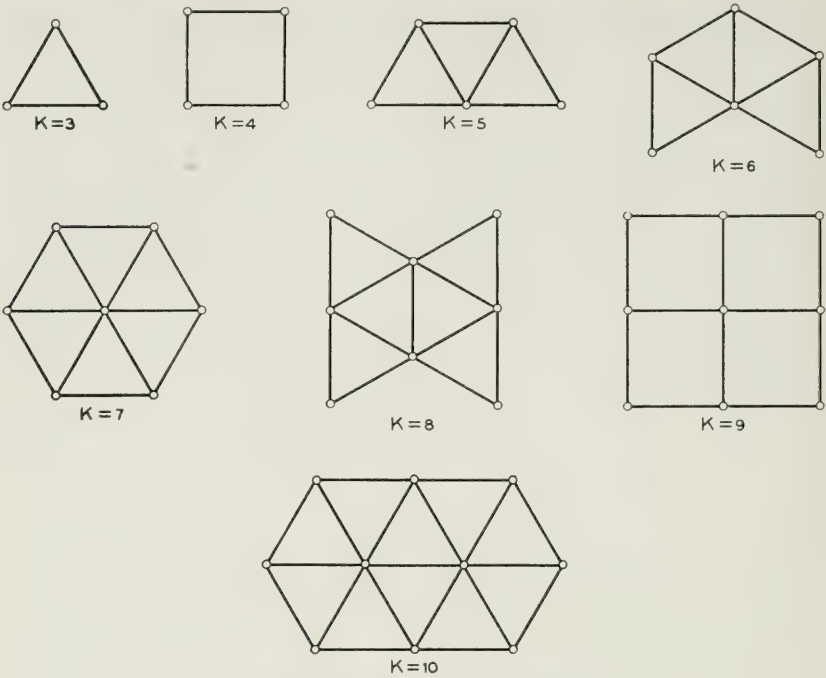


Fig. 3—Two dimensional alphabets.

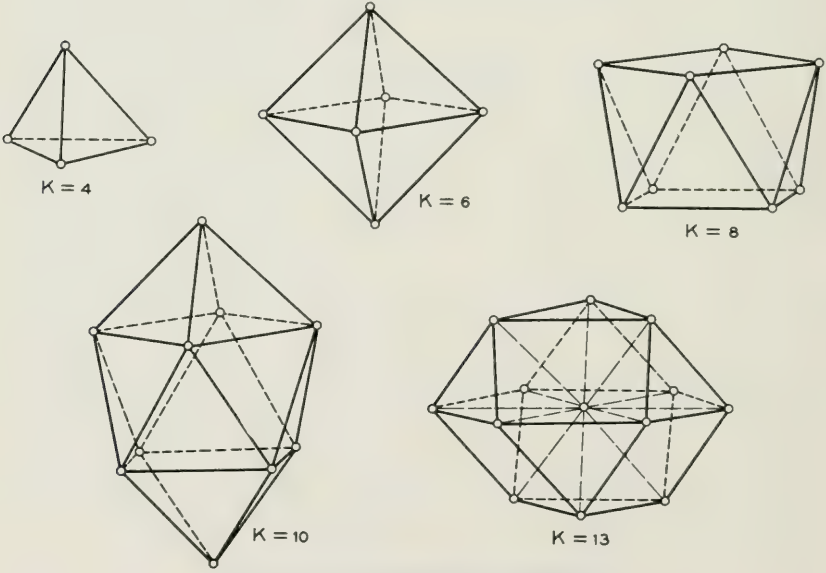


Fig. 4—Three dimensional alphabets.

where r_1 is the rms distance from the origin to the points of a configuration A , R_0 is the distance from the origin to the centroid of A , and r_2 is the rms distance from the points of A to the centroid of A .

In plotting points on the efficiency graph the notation K, D is used for the best K -letter D -dimensional alphabet which has been found. The arrangement of points for various $K, 2$ and $K, 3$ alphabets is given in Figs. 3 and 4. In these figures two points are joined by a straight line if the distance between them is 1 (which is the value we have adopted for the minimum allowed separation $2r_0$). Although not shown, the origin is always at the centroid of the figure. To aid interpretation of these diagrams we have included Fig. 5 which demonstrates how all the signals of a typical alphabet can be generated. The functions of time shown in

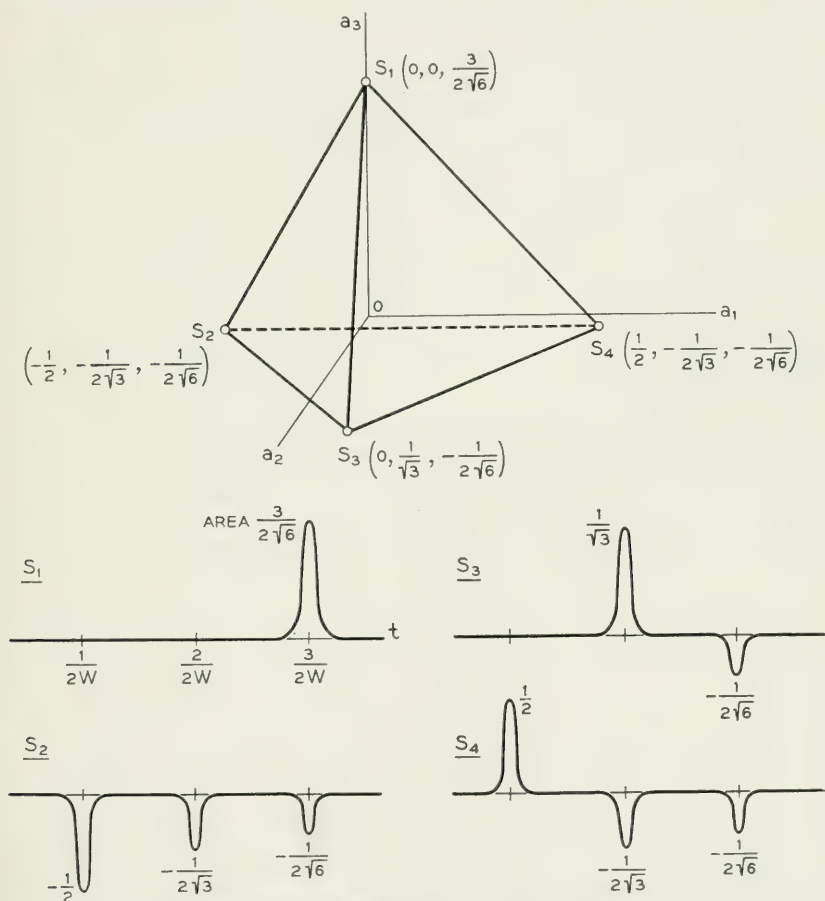


Fig. 5—Generation of the 4,3 code signals.

Fig. 5 are not the code signals themselves but impulse functions which are to be passed through a low pass filter with cutoff at W c.p.s. to form the code signals.

The best possible higher dimensional alphabets can be described more easily verbally than pictorially. In four dimensions we have found four alphabets.

The 25_4 alphabet consists of the origin and all 24 points in 4 dimensional space having two coordinates equal to zero and the remaining two equal to $1/\sqrt{2}$ or $-1/\sqrt{2}$. Each of the 24 points lies a unit distance away from the origin and its 10 other nearest neighbors; they are, in fact, the vertices of a regular solid. This alphabet has an advantage beyond its high efficiency. The code signals are composed entirely of positive and negative pulses of fixed energy and so should be easier to generate than most of the other codes which appear in this paper.

The 800_4 alphabet is constructed in the following way: Consider a lattice of points throughout the entire 4-dimensional space formed by taking all the linear combinations with integer coefficients of a basic set of four vectors. That is, the lattice points are of the form $C_1v_1 + C_2v_2 + C_3v_3 + C_4v_4$ where C_1, \dots, C_4 are integers and the v_i are the four given vectors. In connection with our problem it is of interest to know what lattice, (i.e. what choice of v_1, v_2, v_3, v_4) has all lattice points separated at least unit distance from one another and at the same time packs as many points as possible into the space per unit volume. When a solution to this "packing problem" is known, it is clear that a good alphabet can be obtained just by using all the lattice points which are contained inside a hypersphere about the origin as the letter points. Many of the two dimensional alphabets illustrated in the sketches are related in this way to the corresponding two dimensional packing problem (which is solved by letting v_1 and v_2 be a pair of unit vectors 60° apart). A solution to the four dimensional packing problem is afforded by

$$v_1 = \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0, 0$$

$$v_2 = \frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}}, 0$$

$$v_3 = \frac{1}{\sqrt{2}}, 0, 0, \frac{1}{\sqrt{2}}$$

$$v_4 = 0, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0.$$

This lattice contains two points per unit volume (twice as dense as the cubic lattice in which v_1, \dots, v_4 are orthogonal to one another) and each

point has 18 nearest neighbors. A hypersphere of radius 3 about the origin has a volume $(\pi^2/2)3^4$, about 400. Thus it contains about 800 lattice points. Take these as the code points of the 800, 4 code. Their average squared distances from the origin can be estimated as

$$\frac{\int_0^3 r^5 dr}{\int_0^3 r^3 dr} = \frac{2}{3} (3)^2 = 6.$$

N in Equation (13) may be estimated at 18; this is conservative because some lattice points outside the sphere are being counted.

The two remaining four dimensional alphabets belong to two families of D -dimensional alphabets.

The 4, 3; 5, 4; \dots ; $D + 1, D \dots$ alphabets are the vertices of the simplest regular solid in D -dimensional space. For example, 4, 3 is a tetrahedron. Such a solid can be constructed from $D + 1$ vertices whose coordinates are the first $D + 1$ rows of the scheme

0	0	0	0	0	\dots
1	0	0	0	0	\dots
$\frac{1}{2}$	$\frac{3}{2\sqrt{3}}$	0	0	0	\dots
$\frac{1}{2}$	$\frac{1}{2\sqrt{3}}$	$\frac{4}{2\sqrt{6}}$	0	0	\dots
$\frac{1}{2}$	$\frac{1}{2\sqrt{3}}$	$\frac{1}{2\sqrt{6}}$	$\frac{5}{2\sqrt{10}}$	0	\dots
$\frac{1}{2}$	$\frac{1}{2\sqrt{3}}$	$\frac{1}{2\sqrt{6}}$	$\frac{1}{2\sqrt{10}}$	$\frac{6}{2\sqrt{15}}$	\dots
.	\dots
.	\dots
.	\dots

The vertices all lie a distance $\sqrt{D/2(D + 1)}$ from the centroid of the figure.

6, 3; 8, 4; \dots ; $2D, D, \dots$ are obtained by placing a point wherever any positive or negative coordinate axis intersects the sphere of radius

$1/\sqrt{2}$ about the origin. Thus it follows that 6, 3 consists of the vertices of an octohedron.

Error correcting alphabets ((k, K, D)): The error correcting alphabets discussed in Part I can be converted into good alphabets for this channel by replacing all digits which equalled 0 by -1 . Three error correcting alphabets appear on the chart; each is labelled by three numbers signifying (k, K, D) .

Slepian alphabets (SD): Using group theoretic methods, D. Slepian has attempted to construct families of alphabets which signal at rates approaching C . Although this goal has not yet been reached, families of alphabets depending on the parameter D have been found which approach the ideal curve to within 6.2 db and then get worse as $D \rightarrow \infty$. In the simplest of these families of alphabets, $D = 2m$ is even and the letters consist of all the $2^m C_{2m, m}$ sequences containing m zeros, the remaining places being filled by ± 1 . The best alphabet in this family is the one with $D = 24$. It lies 6.23 db away from the ideal curve and contains 1.1×10^{10} letters. The alphabets of this family for $D = 10, 24$, and 70 appear on the efficiency graph labelled $S10, S24$, and $S70$.

The conclusion to which one is forced as a result of this investigation is that one cannot signal over a channel with signal to noise level much less than 7 db above the ideal level of Equation (2) without using an unbelievably complicated alphabet. No ten digit alphabet tolerates less than 7.7 db more than the ideal signal to noise ratio.

It would be interesting to know more about good higher dimensional alphabets. They are very much more difficult to obtain. The regular solids, which provided some good alphabets in 3 and 4 dimensions, provide nothing new in 5 or more dimensions; there are only three of them and they correspond to our $D + 1, D; 2D, D$, and 2^D binary alphabets. Worse still, the packing problem also becomes unmanageable after dimension 5.

ACKNOWLEDGMENT

The author wishes to thank R. W. Hamming, L. A. MacColl, B. McMillan, C. E. Shannon, and D. Slepian for many helpful suggestions during the investigation summarized by this paper.

Principle Strains in Cable Sheaths and Other Buckled Surfaces

By I. L. HOPKINS

(Manuscript Received February 25, 1952)

Equations are developed for rigorous determination of magnitudes and directions of principal strains in plastic deformation, by means of measurements of rectangular strain rosettes. Application to the study of telephone cable sheath is described.

In the course of certain studies of the polyethylene used in the sheath of telephone cable, it was necessary to calculate the magnitudes and directions of the principal strains from data obtained by measurements of the distortion of a square grid which had previously been printed on the surface of the cable. The strains were large, rendering useless the usual expressions for analysis of strain rosette data¹. Such large strains are characteristically sustained for a wide variety of high polymeric materials of increasing importance for wire and cable sheathing as well as other structural uses. In this article the requisite formulas are developed.

The basic assumptions are:

- (1) The strains may be large.
- (2) The strains are uniform over any square of the grid (equivalent to the condition that a square transforms into a parallelogram).
- (3) The square may be regarded as plane.
- (4) Two of the principal strains are parallel to the surface.

We shall first consider only the two principal strains in the plane of the surface of the cable. Suppose these two strains to be parallel with the x and y coordinate axes, respectively, and that one side of the square is aligned, before straining, at the angle ϕ with the x axis. This is illustrated in Fig. 1.

Let e_x = maximum principal strain

e_y = minimum principal strain

$\lambda_x = 1 + e_x$

$\lambda_y = 1 + e_y$

¹ Cf. for example, Max Frocht, "Photoelasticity," **1**, p. 37, 1941.

If primes are used to refer to the strained state,

$$\lambda_x = \frac{x'_b - x'_a}{x_b - x_a} = \frac{x'_d - x'_c}{x_d - x_c}$$

$$\lambda_y = \frac{y'_b - y'_a}{y_b - y_a} = \frac{y'_d - y'_c}{y_d - y_c}$$

If L_1 and L_2 are the lengths of the sides of the unstrained square, and L_3 and L_4 the diagonals,

$$(L_1 + \Delta L_1)^2 = \lambda_x^2 (x_b - x_a)^2 + \lambda_y^2 (y_b - y_a)^2 \quad (1a)$$

$$(L_2 + \Delta L_2)^2 = \lambda_x^2 (x_d - x_c)^2 + \lambda_y^2 (y_d - y_c)^2 \quad (1b)$$

$$(x_b - x_a)^2 = (y_d - y_c)^2 = L_1^2 \cos^2 \phi_1 = L_2^2 \cos^2 \phi_1 \quad (2a)$$

$$(y_b - y_a)^2 = (x_d - x_c)^2 = L_1^2 \sin^2 \phi_1 = L_2^2 \sin^2 \phi_1 \quad (2b)$$

whence, if

$$\frac{L_1 + \Delta L_1}{L_1} = L'_1, \quad \frac{L_2 + \Delta L_2}{L_2} = L'_2, \text{ etc.} \quad (2c)$$

$$L_1'^2 = \lambda_x^2 \cos^2 \phi_1 + \lambda_y^2 \sin^2 \phi_1 \quad (3a)$$

$$L_2'^2 = \lambda_x^2 \sin^2 \phi_1 + \lambda_y^2 \cos^2 \phi_1 \quad (3b)$$

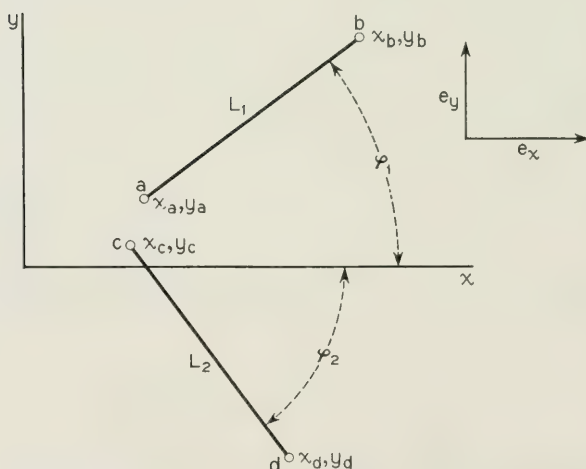


Fig. 1—Lines ab and cd , before the xy plane is strained by stretching (or compressing) in the x and y directions.

Henceforth, for clarity, suppose the subscript "1" to refer to the longer side of the parallelogram, "2" to the shorter side, "3" to the longer diagonal, and "4" to the shorter.

$$S_1 = \text{original slope of } L_1 = \tan \phi_1 = \frac{y_b - y_a}{x_b - x_a}$$

$$S_2 = \text{original slope of } L_2 = \tan \phi_2 = \frac{y_d - y_c}{x_d - x_c}$$

$$S'_1 = \tan \phi'_1 = \frac{\lambda_y}{\lambda_x} S_1 \quad (4a)$$

$$S'_2 = \tan \phi'_2 = \frac{\lambda_y}{\lambda_x} S_2 \quad (4b)$$

Since $\phi_1 - \phi_2 = 90^\circ$,

$$S_2 = -\frac{1}{S_1} \quad (5a)$$

$$S'_2 = -\lambda_y/\lambda_x S_1 \quad (5a)$$

By expansion and substitution from Equations (4) and (5),

$$\tan (\phi'_1 - \phi'_2) = \frac{\frac{\lambda_y}{\lambda_x} \left(S_1 + \frac{1}{S_1} \right)}{1 - \left(\frac{\lambda_y}{\lambda_x} \right)^2} \quad (6)$$

Let

$$\psi = 90^\circ - (\phi'_1 - \phi'_2)$$

then

$$\tan (90^\circ - (\phi'_1 - \phi'_2)) = \tan \psi = \frac{1 - \left(\frac{\lambda_y}{\lambda_x} \right)^2}{\frac{\lambda_y}{\lambda_x} \left(S_1 + \frac{1}{S_1} \right)} \quad (7)$$

which is the shear between L'_1 and L'_2 .

$$(S_1 + 1/S_1) = \tan \phi_1 + \cot \phi_1 = \frac{2}{\sin 2\phi_1} \quad (8)$$

and substituting this in equation (7),

$$\sin 2\phi_1 = \frac{2\lambda_x\lambda_y \tan \psi}{(\lambda_x^2 - \lambda_y^2)} \quad (9)$$

whence

$$\cos 2\phi_1 = \sqrt{1 - \frac{4\lambda_x^2 \lambda_y^2 \tan^2 \psi}{(\lambda_x^2 - \lambda_y^2)^2}} \quad (10)$$

Remembering that

$$\cos^2 \phi_1 = \frac{1 + \cos 2\phi_1}{2} \quad (11a)$$

and

$$\sin^2 \phi_1 = \frac{1 - \cos 2\phi_1}{2} \quad (11b)$$

and substituting Equation (10) in Equation (11), Equation (11) in Equation (3), and then solving the quadratic equation thus formed for λ_x and λ_y , we have

$$\lambda_x^2, \lambda_y^2 = \frac{(L_1'^2 + L_2'^2) \pm \sqrt{(L_1'^2 + L_2'^2)^2 - 4L_1'^2 L_2'^2 \cos^2 \psi}}{2} \quad (12)$$

Referring to Fig. 2, and using the law of cosines, and remembering that L_3' is the ratio of the strained to the unstrained length of the diagonal,

$$-\cos \theta = \sin \psi = \frac{2L_3'^2 - (L_1'^2 + L_2'^2)}{2L_1' L_2'} \quad (13a)$$

whence

$$\cos^2 \psi = \frac{4L_3'^2 (L_1'^2 + L_2'^2 - L_3'^2) - (L_1'^2 - L_2'^2)^2}{4L_1'^2 L_2'^2} \quad (13b)$$

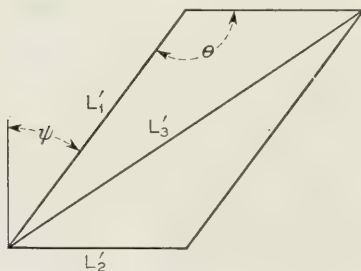


Fig. 2—A parallelogram formed by straining a square. L_1' , L_2' and L_3' are the ratios of the lengths of the indicated lines to their original lengths.

This expression, substituted in Equation (12) and reduced, gives

$$\lambda_x^2, \lambda_y^2 = \frac{(L_1'^2 + L_2'^2) \pm \sqrt{2(L_1'^2 - L_3'^2)^2 + 2(L_2'^2 - L_3'^2)^2}}{2} \quad (14)$$

It may be noted here that a property of the parallelogram, namely, in the notation used here,

$$L_1'^2 + L_2'^2 = L_3'^2 + L_4'^2 \quad (15)$$

makes it immaterial which diagonal is used. This may be readily seen by substituting

$$L_3'^2 = L_1'^2 + L_2'^2 - L_4'^2$$

in Equation (14). The effect is merely that of substituting L_4 for L_3 . In Equation (13a), however, the result is a change in the sign of ψ .

As an example of the application of these equations, the measurements of one specimen were:

$$L_1' = 2.1$$

$$L_2' = 1.2$$

$$L_3' = 2.0$$

From Equation (14),

$$\lambda_x^2 = 4.758, \quad \lambda_x = 2.181, \quad e_x = 1.181$$

$$\lambda_y^2 = 1.092 \quad \lambda_y = 1.045 \quad e_y = 0.045$$

From Equation (13a),

$$\sin \psi = 0.4266, \text{ whence}$$

$$\psi = 25.3^\circ$$

$$\tan \psi = 0.472$$

From Equation (9),

$$\sin 2\phi_1 = 0.587$$

$$\phi_1 = 18.0^\circ$$

$$\tan \phi_1 = 0.324$$

From Equation (4a),

$$\tan \phi_1' = 0.1554$$

$$\phi_1' = 8.83^\circ$$

From Equation (9), it is obvious that the maximum value of $\tan \psi$ occurs at $\phi_1 = 45^\circ$, and is in this case equal to 0.804.

This example is illustrated in Fig. 3.

The question of direction of the x and y axes is simply settled by drawing a line through either of the acute angles of the parallelogram, crossing the parallelogram at an angle ϕ'_1 with the longer side. This line will be parallel to the x direction, which is, according to the convention, that of greatest strain.

So far no mention has been made of strain in the third dimension; that is, a change in thickness of the sheath. In plastic deformation, the volume change is generally negligible. This requires that

$$\lambda_x \lambda_y \lambda_z = 1$$

whence

$$\lambda_z = \frac{1}{\lambda_x \lambda_y}$$

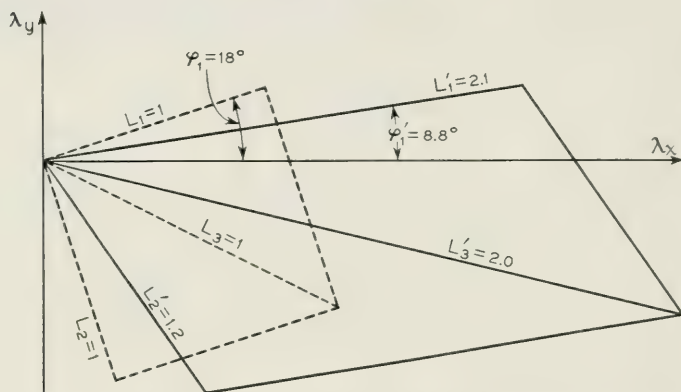


Fig. 3—A square and the parallelogram resulting from stretching to length ratios $\lambda_x = 2.181$ in the x -direction and $\lambda_y = 1.045$ in the y -direction.

TABLE I

Degrees of twist in 3 feet	Principal Strains, per cent		
	Parallel to Surface		Perpendicular to Surface
	Max.	Min.	
180	16	06	-19
270	26	09	-27
360	33	14	-34
450	36	20	-39
540	42	19	-41
630	43	22	-43
720	46	24	-45

In the example given,

$$\lambda_z = 0.439, \quad c_z = -0.561$$

Polyethylene sheaths of cable specimens 3 feet long buckled severely over their entire length when the cables were twisted 720° and showed the strains given in Table I at steps up to the final twist².

The ratio of maximum to minimum strain parallel to the surface is about 2:1. Tests with a 1:1 ratio³, a more severe condition, have shown that the principal strains at rupture will be of the order of 300 per cent. Therefore it is evident that the strains incidental to the most severe types of handling will not, of themselves, cause rupture of the sheaths.

² Unpublished memorandum by A. G. Hall.

³ I. L. Hopkins, W. O. Baker and J. B. Howard, *J. Appl. Phys.*, **21**, No. 3, pp. 206-213, March, 1950.

A New Recording Medium For Transcribed Message Services

By JAMES Z. MENARD

(Manuscript received March 10, 1952)

A magnetic recording medium composed of rubber impregnated with magnetic oxide and lubricant is particularly suited to applications requiring the continuous repetition of short transcribed messages. It affords exceptional life, reliability, and economy in telephone applications, where it is utilized in the form of molded bands stretched over cylinders of the recording mechanisms.

In the Bell System there are several applications requiring the repetition of short voice announcements. Some of the existing applications are weather announcements, intercept of calls to vacant and unassigned numbers, quotations of delays on long-distance calls, and certain leased industrial services, such as stock price quotation. Most of these require continuous repetition of messages between 5 and 60 seconds in length. In some the message remains fixed but in others it is changed at frequent intervals.

Magnetic recorders offer particular advantages for services of this nature, because they require a minimum of equipment and operating skill to produce durable records which are instantly reproducible without processing. For several years the Bell System has used a magnetic recorder employing a loop of Vicalloy tape in the 3A announcement system to furnish weather announcements, and a similar type of recorder has been used in a leased industrial system at the *New York Times*.

Recently these Laboratories have undertaken the development of transcribed message facilities to meet additional service applications. It was required that the new facilities should provide satisfactory transmission quality and afford considerable flexibility in regard to message length, but the paramount requirement was for reliability and long life.

It did not appear practicable to extend the techniques of the Vicalloy tape machine to give the flexibility, convenience of operation and re-

liability desired in the new applications, and attention was therefore directed to two new classes of magnetic recording media which have been developed in recent years. These are the electroplated media and the powdered media.

In recent years magnetic recording media have been commercially produced by an electroplating process by the Brush Development Company of Cleveland, Ohio. Evaluation by Bell Telephone Laboratories shows that such a plating does not easily deteriorate, gives a relatively high signal output and is capable of excellent transmission characteristics. But in order to realize consistently satisfactory transmission, it is necessary to maintain intimate contact between the recording medium and the magnetic recording and reproducing heads. The expense of providing the relatively precise mechanisms necessary to obtain the desired performance objectives suggested the exploration of other media which might simplify this problem.

The powdered magnetic media have evolved from German work dating back to about 1932 and from American work since about 1941. In these media the active magnetic material is a finely divided ferro-magnetic powder, usually iron oxide. This is usually applied with a binder as a surface coating on a tape of plastic or paper, but the Germans at one time produced a tape which was a homogeneous mixture of oxide and plastic. In their present state of development, media of this type offer excellent transmission characteristics and are relatively economical. In the past four or five years they have found widespread commercial application in the form of coated tape in all fields of recording and transcription work.

Attempts were made to employ commercial types of these coated tapes in various forms of continuous-loop mechanisms, but none met the desired requirements in regard to life, reliability, and flexibility of operation. An analysis of the experimental results indicated that most failures were due to physical failure of the media as a result of the tension, flexion and abrasion to which they were subjected, but the magnetic records were substantially undeteriorated even when physical failure of the supporting base occurred.

It became apparent that a specialized recording medium would have to be developed to meet the Bell System requirements for transcribed message services. Development effort was concentrated on the field of powdered media, because these media offered attractive transmission properties and because the expanding commercial importance of this field promised a continuing industrial development and production program which would provide an economical source of high quality magnetic

materials. The following premises guided the development program:

(1) The recording medium should be subjected to the least possible physical manipulation in use to minimize failures. To accomplish this it was decided to develop the recording medium in a form suitable for use on the surface of a rotating cylinder and to use a helical recording track on this surface when the message length required more than one revolution of the cylinder. It was hoped that with this arrangement, physical failure of the recording medium would be eliminated, and the service life would be determined by the wear occurring between the medium and the magnetic pole pieces.

(2) The recording medium should exhibit some compliance to facilitate intimate contact with the magnetic pole-pieces.

(3) The transmission quality should meet present-day telephone standards for transmission of speech. The higher quality necessary for the recording of music, while desirable, should not be considered a requirement.

A number of experimental powdered media were prepared and tested. These all utilized commercially available iron oxide powder with a coercive force of approximately 250 oersteds, and the samples included coated media, made by dipping, spraying and doctoring the coating on various base materials, and impregnated media, prepared by mixing the oxide in the base material and forming the mixture.

A medium consisting primarily of an elastic rubber band impregnated with magnetic particles was found to be particularly suited to applications requiring long life in continuous service. A study of compounding and manufacturing processes for this medium was made by the rubber products group at these Laboratories under the direction of H. Peters, and the compound which evolved consists primarily of synthetic rubber loaded with magnetic iron oxide, and containing small amounts of lubricants, inhibitors and curing agents. The compound is decidedly rubber-like in character, and is utilized in the form of seamless bands about $\frac{1}{32}$ to $\frac{1}{8}$ inches thick, which are stretched over the surface of cylinders about 10 per cent larger than the bands.

The bands are prepared by thoroughly milling together the following:

- 100 Parts by weight type GN neoprene
- 100 Parts by weight magnetic iron oxide
- 5 Parts by weight zinc oxide
- 4 Parts by weight magnesium oxide
- 2 Parts by weight paraffin

and forming the compound into bands by conventional rubber molding and curing techniques. The resulting bands show a tensile strength of about 2500 pounds per square inch, and the elongation before breaking is about 700 per cent. No particularly difficult manufacturing problems are encountered, and present evidence indicates that satisfactory overall quality control can be achieved by carefully controlling the compounding constituents, the milling and the molding.

Several bands which are used in telephone services are shown in Fig. 1. These bands are utilized in recorder-reproducer mechanisms by stretching them over a cylinder, on which pivoted magnetic pole-pieces trace a cylindrical or a helical track as it rotates.

When the bands are first taken from the mold they exhibit a high coefficient of friction. After a few hours enough paraffin migrates to the surface to form a thin, slippery film. If the bands are then put into service the pole-pieces form a polished track and the continuing migration of paraffin maintains the lubricating film between the band and the pole-pieces.

If this recording medium is used intermittently, the self-lubrication may cause difficulty. The migration of lubricant to the recording surface is continuous, and the lubricant may accumulate on the surface in sufficient thickness to impair the contact with the magnetic head if the recording equipment is not operated for several weeks. It may then be

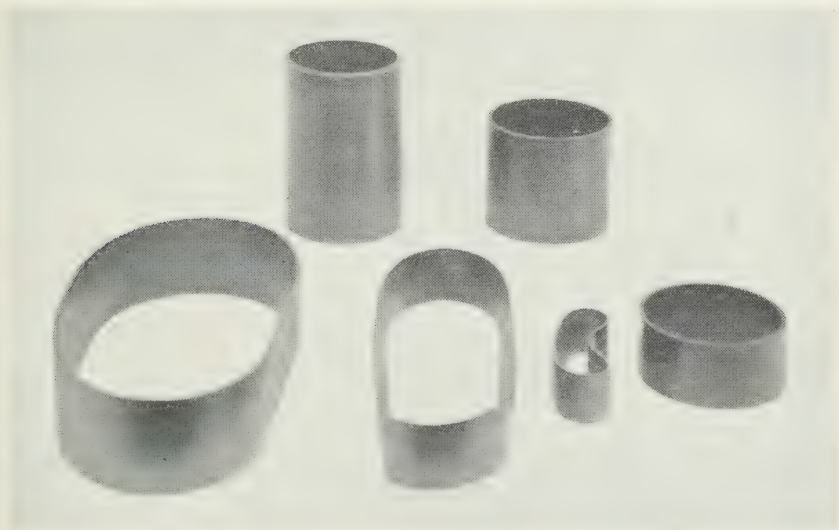


Fig. 1—Typical magnetic rubber bands used in telephone applications.

necessary to wipe the excess lubricant from the surface to obtain satisfactory operation.

The lubricant in this particular recording medium was chosen for operation in central offices and similar locations where temperature ranges are moderate. If extreme temperatures are to be encountered the lubricant problem will have to be re-examined. Continued research in this field should result in improvement in this characteristic.

In life tests, five million message repetitions have been obtained with insignificant wear of the band and the magnetic head, and with no measurable deterioration in the level and quality of the recording after an initial level drop of about 2 db which occurs during the first few reproductions. The head pressure is a significant factor affecting the life, and in these tests a head pressure of 25 grams was used with a 0.100 inch wide head.

This medium represents some compromise in the attainable transmission properties to favor the physical properties desired for reliability and long life, but the transmission is entirely adequate for the intended applications.

A typical frequency response characteristic is shown in Fig. 2. This is representative of the results obtained when the equipment is maintained by field personnel. The output level, also indicated by Fig. 2, is from 8 to 12 db below that obtained from commercial coated magnetic tape. This is largely because the concentration of magnetic oxide, on a volume basis, cannot be made as high in the impregnated material as is possible in the coating of conventional tape. This is not a serious disadvantage, however, as the level is high enough to permit amplification without special precautions in regard to noise.

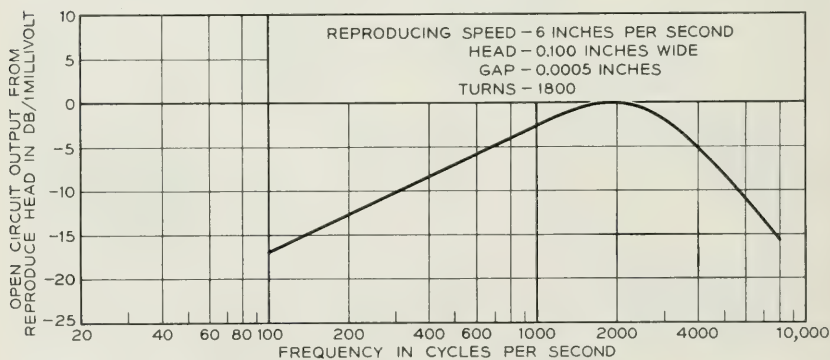


Fig. 2—Frequency response of magnetic recording equipment using iron oxide impregnated molded neoprene bands.

When ring type magnetic heads are used for recording, these bands exhibit frequency response characteristics quite similar to coated tapes using the same magnetic oxides, although the bands are of homogeneous magnetic material up to $\frac{1}{8}$ inch thick and the tapes have magnetic coatings less than 0.001 inches thick. This is because the field from the recording gap becomes ineffective at a distance of about 0.001 inches, and the signal is recorded only on a thin surface layer of the medium, regardless of its total thickness.

The noise characteristic of this medium is somewhat unusual. It has been shown* that the reproducing process is not restricted to the surface layer of the medium, but that to a first approximation, when the medium has low permeability, the signal from a magnetized element at any depth in the recording medium will be attenuated with respect to the signal produced by the same element in intimate contact with the reproducing head by the factor:

$$\frac{55 \text{ decibels} \times S}{\lambda}$$

where S = distance between magnet and head

λ = "wavelength" of magnet

This indicates, for example, that the signal from a magnetized element at a depth of $\lambda/2.75$ will be attenuated by only about 20 db and may therefore make significant contribution to the total output.

In the Bell System telephone applications, where a transmission bandwidth of 100 to 4000 cycles per second is required, the belts are run at a speed of about 6 inches per second. The wavelength at 100 cycles per second is then 0.060 inches, and at this frequency significant output can be obtained from a layer about 0.02 inches thick. The desired recording is limited to a layer about 0.001 inches thick, but a layer of about twenty times this thickness may contribute to noise. As a consequence, at low frequencies this medium tends to exhibit higher background noise than do the coated tapes. The magnitude of the noise is appreciably affected by the method of erasure.

Two methods of erasing a magnetic record are known to the art. These are the saturation erase, in which the magnetic record is exposed to a unidirectional magnetic field of saturation intensity, and the neutralization erase in which the magnetic record is exposed to an alternating field which reaches saturation intensity and decreases cyclically to zero

* R. L. Wallace, Jr., "Reproduction of Magnetically Recorded Signals," *Bell System Tech. J.*, Oct., 1951.

over a period of several cycles. It is well known that a neutralization erase results in a residual background noise which may be as much as an order of magnitude below that produced by a saturation erase. The neutralization erase is therefore widely used in tape recording, and is obtained by energizing the erase head with alternating current of a frequency several times the highest signal frequency passed by the recording equipment.

With the impregnated recording medium, the recorded signal can be successfully erased by using a conventional ring-type erase head energized with high-frequency current. The field from this type of erasing effectively neutralizes the surface layer which contains the recorded signal, but does not penetrate appreciably beyond. Therefore, if precautions are not observed, the lower layers of this medium beyond the reach of the erase field may acquire a random cumulative magnetization from switching surges, accidental exposure to magnetized tools and strong fields, and this will be evidenced by a gradual deterioration in the signal to noise ratio at the low-frequency end of the transmission band. The quality, however, remains entirely adequate for commercial telephone use.

The foregoing limitations are minimized by an erasing method which has been developed at these Laboratories for applications where it is convenient to erase the entire message in one revolution of the recording cylinder, preparatory to recording a new message. This method employs an erasing structure in the form of an E-shaped stack of magnetic laminations, carrying on the center leg a coil which is energized by low-frequency (60 cycle) alternating current. The lamination stack is approximately the width of the recording medium, and the gaps between the center leg and each side leg are about $\frac{1}{4}$ inch wide. When this structure is spaced about $\frac{1}{16}$ inch from the surface of the recording medium traveling at 6 inches per second or less, and is energized by 60-cycle power to produce a maximum field of about 2000 gauss, the entire thickness of the recording medium is subjected to an alternating magnetic field which reaches saturation intensity and over a period of several cycles decreases progressively to zero. This effectively demagnetizes the full thickness of the recording medium. If the current is switched off with the erase structure in operating position, those elements of magnetic material within the field at that instant would be subjected to no further reversals and would consequently behave substantially as if they had been subjected to a direct-current magnetic field of the same intensity as the alternating-current field at the time it was interrupted. The section of record medium under the influence of the erase structure at the time it was de-energized would exhibit excessive noise in com-

parison with the remainder of the record-medium which was subjected to the normal alternating-current erase. This effect becomes negligible if the separation between the record medium and the erase structure is increased by $\frac{1}{2}$ inch before the current is interrupted. This is accomplished by using a solenoid-actuated mounting for the erase structure so arranged that the structure normally is retracted from the erasing position and holds open a switch in the circuit to its coil. When erasure is desired, the solenoid is actuated. This moves the erase structure into operating position and releases the switch to energize the coil. When the erase cycle is completed the solenoid is de-energized, and the erase structure retracts, opening the switch at the end of its travel. Fig. 3 is a sketch

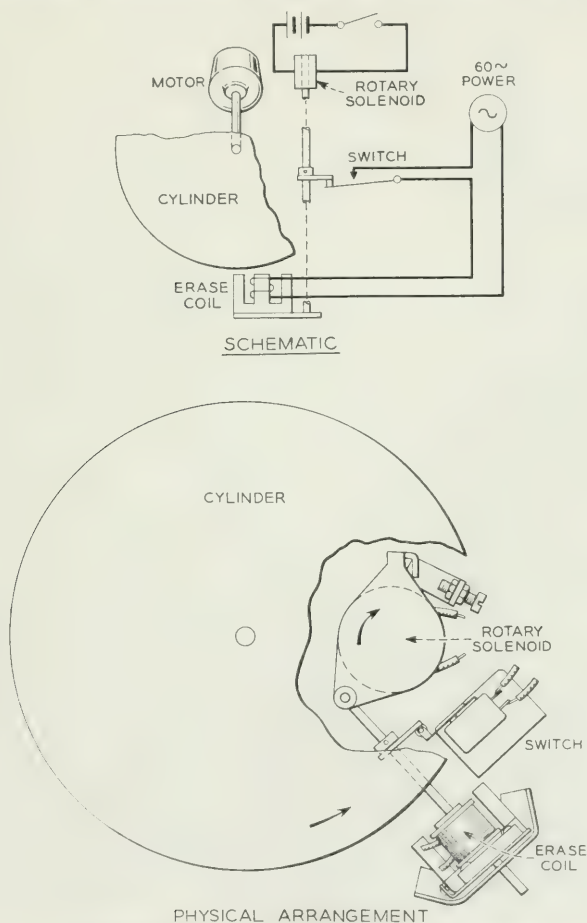


Fig. 3—Method of erasing magnetic recorder.

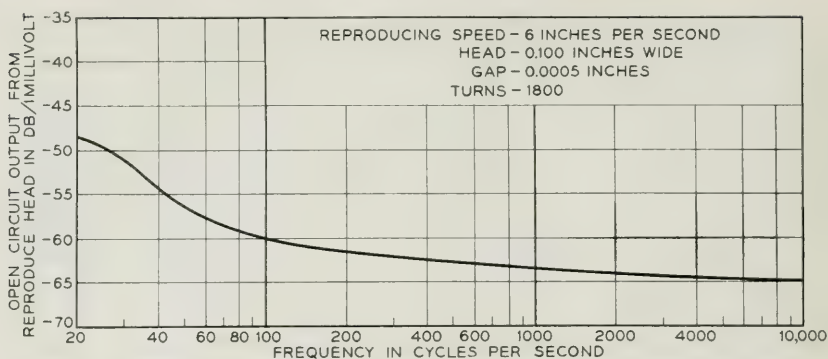


Fig. 4—Typical noise spectrum of $\frac{1}{8}$ -inch iron oxide impregnated molded neoprene bands measured in 200 cycle bands after neutralization erase with 60-cycle ac field.

showing the application of this erase method to a cylinder-type machine. This method of erase results in a background noise level measured unweighted over a 4000-cycle band which is at least 45 db below a 1000-cycle signal recorded with 4 per cent total distortion. A typical background noise spectrum is shown in Fig. 4.

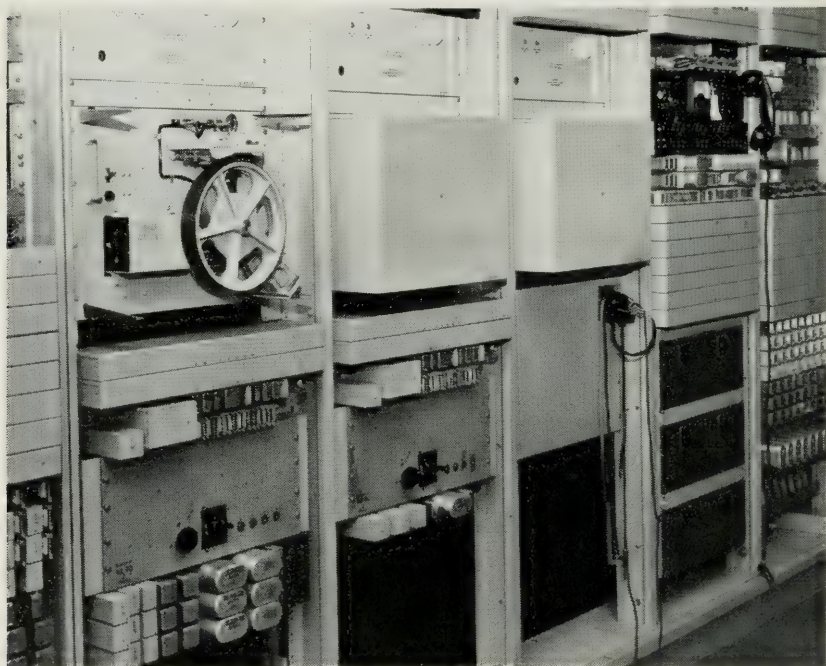


Fig. 5—General view of recording machines in 3A announcement system at Cleveland.

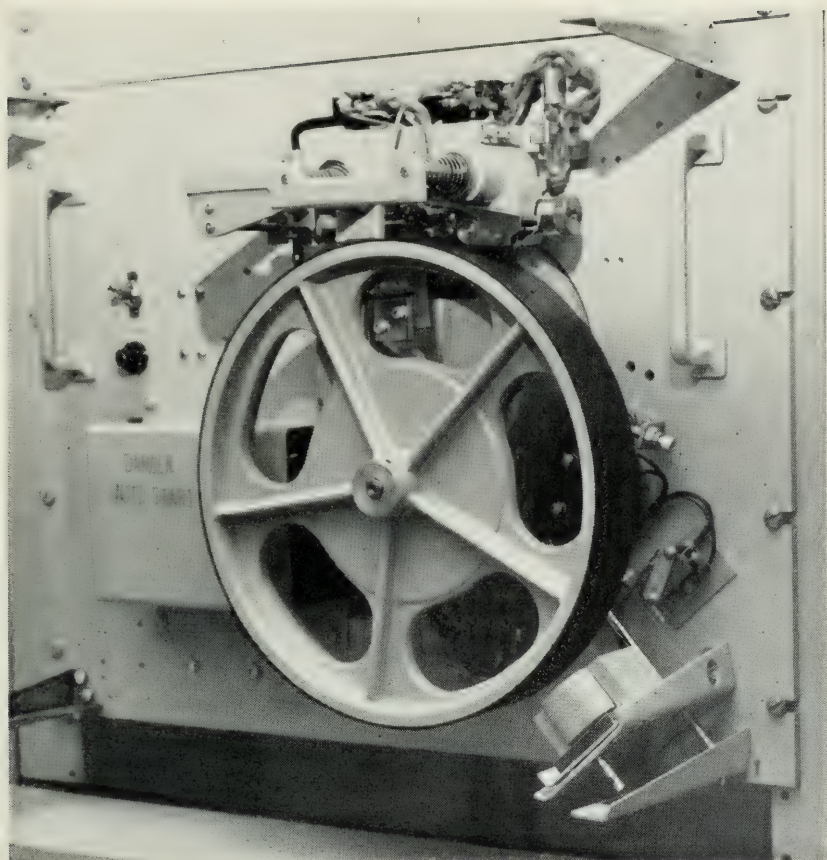


Fig. 6—Closeup of recording machine in 3A announcement system at Cleveland.

The first installation of transcribed message equipment employing this new medium was in the 3A announcement system at Cleveland, Ohio, to supply weather announcement service.

The magnetic recording equipment in this installation is a cylinder-type mechanism with associated recording-reproducing amplifier. The mechanism uses bands $\frac{1}{16}$ inch thick, $1\frac{5}{8}$ inches wide and $7\frac{1}{16}$ inches in diameter, stretched over a cylinder 9 inches in diameter. A single record-reproduce head in a pivoted mounting is cam-controlled to trace a helix on the cylinder. The cam is coupled to the cylinder via a quick-change gear train which gives a choice of a three-turn, a five-turn or an eight-turn helix, and the cylinder is driven from a gear-reducer which allows a choice of two slightly different operating speeds. Six different cycle times, ranging from about ten seconds to about 45 seconds, are provided

by the two operating speeds and the three cam ratios. Approximately 90 per cent of any cycle time is available for recording or reproduction, and the remaining 10 per cent is occupied by the return of the head to the beginning of the helix.

The recorded track is 0.100 inches wide, and when an eight-turn helix is used, there is a separation of 0.025 inches between tracks. The previously described low-frequency alternating current erase is used.

The 3A announcement system employs three channels of this recording equipment in a complex control circuit which provides facilities for erasing, recording, monitoring and automatic switchover to stand-by channels in event of failure. Figs. 5 and 6 show the recording equipment in the Cleveland installation.

Other equipments using this recording medium have been designed to furnish transcribed message service for intercept of calls to vacant, changed and unassigned numbers, to quote delays on long distance calls, and to furnish stock quotation service. Some of these equipments are now undergoing service trials preparatory to standardization for Bell System use.

This new recording medium has been developed to provide the maximum attainable life and reliability in applications requiring an enormous number of repetitions of voice messages. Equipment for such applications is usually located in central offices where the temperature range and other operating conditions are fairly well stabilized. These favorable conditions have facilitated the development of a recording medium which has made it possible to design simple and economical magnetic recorders which are sufficiently versatile and reliable to stimulate the use of transcribed message services to an extent hitherto unrealizable.

There are a number of potential Bell System applications for transcribed message services which do not require an extreme number of message repetitions, but put a premium on low initial cost and trouble-free operation in intermittent service under wide extremes of environment. It may prove desirable to meet the life requirements for applications of this type with a different approach to the lubrication problem, with an unlubricated compound, or with a coated medium which would have some transmission advantages. It is expected that further work in these fields will produce improved recording media for applications of this nature, to expand the field of use in the telephone plant.

Introduction to Formal Realizability Theory—II

By BROCKWAY McMILLAN

(Manuscript received February 15, 1952)

This part of the paper exhibits a network to realize a given positive real impedance matrix.

I. INTRODUCTION TO PART II

1.0 In this part of the paper we prove the following theorem:

1.1 *Theorem:* Let $Z(p)$ be an $n \times n$ matrix whose elements are $Z_{rs}(p)$, $1 \leq r, s \leq n$, where

(i) Each $Z_{rs}(p)$ is a rational function

(ii) $Z_{rs}(\bar{p}) = Z_{rs}(\bar{p})$

(iii) $Z_{rs}(p) = Z_{sr}(p)$

(iv) For each set of real constants k_1, \dots, k_n , the function

$$\varphi_Z(p) = \sum_{r,s=1}^n Z_{rs}(p)k_rk_s$$

has a non-negative real part whenever $\operatorname{Re}(p) > 0$.

Then there exists a finite passive network, a $2n$ -pole, which has the impedance matrix $Z(p)$. A dual result holds for admittance matrices $Y(p)$.

1.2 The converse of this theorem was proved in Part I: that if a finite passive $2n$ -pole has an impedance matrix $Z(p)$, then this matrix has properties (i) through (iv) of 1.1.

1.3 We recall that in Part I matrices satisfying the conditions of 1.1 were called *positive real* (PR).

1.4 The proof of 1.1 is a direct generalization to matrices of the Brune process² for realizing a two-pole impedance function $f(p)$. For this proof we shall require from Part I certain specific properties of positive real operators and matrices. These will be summarized in Section 2 below. Further than this, the present part is almost independent of Part I,

although in terminology, notation, and method a direct continuation of it. References to sections or paragraphs in Part I will be made thus: (I, 6) or (I, 6.23).

1.5 The distinction emphasized in Part I between operators, as abstract geometrical objects, and matrices as concrete arrays of numbers representing these geometrical objects, is not one which we have now to maintain with any strictness. We shall generally preserve it verbally but not use the bracket notation for matrices introduced in Part I.

II. PROPERTIES OF POSITIVE REAL OPERATORS AND MATRICES

2.0 We recall that an impedance operator $Z(p)$ is a linear function from the linear space \mathbf{K} of current vectors k to the linear space \mathbf{V} of voltage vectors v . A positive real operator $Z(p)$ is one whose matrix in any real coordinate frame is positive real. In Section 16 of Part I the following properties of a PR operator $Z(p)$ were established:

2.01 $Z(p)$ has no poles in Γ_+ .*

2.02 If $Re(Z(p)k, k) = 0$ for some $p \in \Gamma_+$, then $Z(p)k \equiv 0$ for all p .

2.03 If it exists, $Z^{-1}(p) = Y(p)$ is PR.

2.04 If $Z(p)$ has a pole at $p = p_0$, it has one at $p = \bar{p}_0$.

2.05 If $Z(p)$ has a pole at $p = i\omega_0$, that pole is simple and

$$Z(p) = \frac{2p}{p^2 + \omega_0^2} R + Z_1(p),$$

where R is real, symmetric, semidefinite, and not zero, and $Z_1(p)$ is PR.

2.06 If $Z(p)$ has a pole at $p = \infty$, that pole is simple and

$$Z(p) = pR + Z_1(p)$$

where R and $Z_1(p)$ are as in 2.05.

2.07 It was emphasized at several points in Part I that the fact of possessing an impedance matrix, and that of possessing an admittance matrix, are each restrictions on a $2n$ -pole \mathbf{N} . It is readily verified from (I, 6.3) and (I, 6.31)—and, indeed, well known—that if \mathbf{N} has both an impedance matrix $Z(p)$ and an admittance matrix $Y(p)$, then

$$Y(p) = Z^{-1}(p).$$

* Γ_+ is the open right half plane: all finite p such that $Re(p) > 0$.

That is, if the impedance matrix of a $2n$ -pole \mathbf{N} is non-singular, then its admittance matrix exists, and conversely.

2.08 It was proved by Cauer⁵, and in (I, 16.8), that if $Z(p)$ is PR and of rank $m < n$, then there exists a real, constant, non-singular matrix W such that

$$Z(p) = W'Z^B(p)W \quad (1)$$

where $Z^B(p)$ is a non-singular $m \times m$ PR matrix bordered by zeros.

2.09 Properties (i) through (iv) of 1.1 define the PR property for a matrix $Z(p)$. A convenient equivalent definition is that

- (i) $Z(p)$ is symmetric,
- (ii) For each $k \in \mathbf{K}$, the function

$$\varphi(p) = (Z(p)k, k)$$

is a PR function of p .

This equivalent definition was established in (I, 16.13).

2.1 In Section 16 of Part I it was also mentioned that there exists for any rational operator $Z(p)$ (PR or not) a numerical function $\delta(Z)$ which generalizes to operators the usual definition of the degree of a rational function. We list here the properties of this degree $\delta(Z)$. They will be established in Section 7.

2.11 $\delta(Z)$ is an integer ≥ 0 .

2.12 If $\delta(Z) = 0$, then $Z(p)$ is a constant—that is, does not depend upon p .

2.13 If $Z^{-1}(p)$ exists, then $\delta(Z) = \delta(Z^{-1})$.

2.14 If $Z(p) = Z_1(p) + Z_2(p)$, where $Z_1(p)$ is finite at every pole of $Z_2(p)$, and $Z_2(p)$ is finite at every pole of $Z_1(p)$, then

$$\delta(Z) = \delta(Z_1) + \delta(Z_2).$$

2.15 If $Z(p) = f(p)R$, where $f(p)$ is a scalar and R is a constant operator, then

$$\delta(Z) = [\text{degree of } f] \cdot [\text{rank of } R].$$

Here the degree of f is the sum

$$\sum_{p_0} [\text{order of the pole of } f(p) \text{ at } p_0]$$

where p_0 runs over all poles of $f(p)$, including ∞ .

2.16 If A and B are constant non-singular matrices, then

$$\delta(Z) = \delta(AZB).$$

It is evident then that $\delta(Z)$ is a geometrical property, being constant over the usual equivalence classes

$$W'Z(p)W$$

or

$$W^{-1}Z(p)W$$

of matrices. Hence we may speak of the degree $\delta(Z)$ of an operator $Z(p)$.

2.17 If $Z(p)$ is formed from an $m \times m$ matrix $Z_1(p)$ by bordering the latter with zeros, then

$$\delta(Z) = \delta(Z_1).$$

2.18 Concerning the degree $\delta(Z)$ we here state a fundamental theorem:

Theorem: The $2n$ -pole whose existence is asserted by 1.1 can be constructed with $\delta(Z)$ reactive elements, and no fewer.

The proof of this theorem will be distributed through Sections 4 and 6. In fact, we must even define exactly the phrase "can be constructed with x reactive elements." This will be done in Section 3.

2.2 *Lemma:* If $Z_1(p)$ and $Z_2(p)$ are PR operators or matrices, then

$$Z(p) = Z_1(p) + Z_2(p)$$

is also PR. If either of $Z_1(p)$ or $Z_2(p)$ is non-singular, then $Z(p)$ is.

Proof: Clearly $Z(p)$ is symmetric. By 2.09, therefore, $Z(p)$ is PR if the function

$$(Z(p)k, k) = (Z_1(p)k, k) + (Z_2(p)k, k) \quad (1)$$

is PR for each $k \in \mathbf{K}$. The right hand side is obviously PR by hypothesis.

If either of $Z_i(p)$ is non-singular, the function (1) cannot vanish in Γ_+ unless $k = 0$ (this is 2.02). Hence in this case $Z(p)$ also is non-singular.

2.21 Clearly 2.2 is independent of the implication, tacit in the notation, that the operators involved are impedances. The lemma holds for PR operators, whether interpreted as operating from \mathbf{K} to \mathbf{V} (impedances) or from \mathbf{V} to \mathbf{K} (admittances).

2.3 In (I, 6.21) and (I, 6.3) it was noted that any $n \times n$ impedance matrix $Z(p)$ defines by fiat a general $2n$ -pole \mathbf{N} whose impedance matrix is that $Z(p)$. Such is the generality of the notion of general $2n$ -pole (I, 4).

Given $2n$ -poles \mathbf{N}_1 and \mathbf{N}_2 , with impedance matrices respectively $Z_1(p)$ and $Z_2(p)$, we know then that there is a general $2n$ -pole \mathbf{N} whose impedance matrix is

$$Z(p) = Z_1(p) + Z_2(p).$$

We call this \mathbf{N} the series combination of \mathbf{N}_1 and \mathbf{N}_2 .

2.31 Designate the terminal pairs of \mathbf{N}_1 by (S_r, S'_r) , those of \mathbf{N}_2 by (T_r, T'_r) , $1 \leq r \leq n$. It is evident that if \mathbf{N}_1 and \mathbf{N}_2 appear in a diagram so connected that

(i) S'_r is connected to T_r , $1 \leq r \leq n$;

(ii) No other connections are made to these nodes;

then the device with terminals S_r, T'_r is \mathbf{N} . This follows at once from Kirchhoff's laws applied to the ideal graph (I, 4.11).

2.32 Dually, if \mathbf{N}_1 and \mathbf{N}_2 have admittance matrices $Y_1(p)$, $Y_2(p)$, then

$$Y(p) = Y_1(p) + Y_2(p)$$

is the matrix of a $2n$ -pole \mathbf{N} defined as the parallel connection of \mathbf{N}_1 and \mathbf{N}_2 . \mathbf{N} is the device whose terminal pairs are formed by joining S_r, T_r and also S'_r, T'_r , $1 \leq r \leq n$.

2.33 Fig. 1 shows the conventions to be used in indicating $2n$ -poles ($n = 4$ in the Figure) with, respectively, impedance matrices and admittance matrices. Fig. 2 then shows the series connection of two impedance devices and the parallel connection of two admittance devices. In each case the terminals on the left are those of the composite device.

2.4 The series and parallel connections just described are special ways of combining $2n$ -poles needed for the generalized Brune process for matrices. They have been introduced here on their merits, as new op-

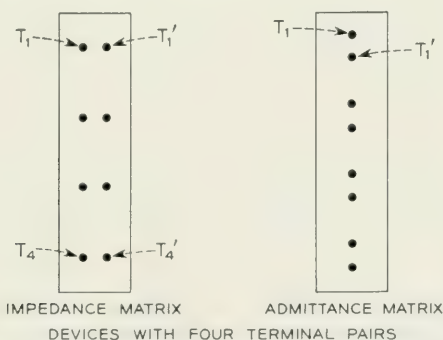


Fig. 1—Conventions used in representing $2N$ poles.

erations. They are, however, expressible in terms of the basic operations of juxtaposition (I, 17) and restriction (I, 18).

For example, the series connection of \mathbf{N}_1 and \mathbf{N}_2 is formed by first juxtaposing \mathbf{N}_1 and \mathbf{N}_2 , to get a $2 \times 2n$ -pole $\bar{\mathbf{N}}$. Let \mathbf{J} be the $2n$ dimensional space of $2n$ -tuples

$$j = [j_1, \dots, j_n, \ell_1, \dots, \ell_n].$$

We interpret this j as a $2n$ -tuple of currents in the $2 \times 2n$ -pole $\bar{\mathbf{N}}$, where j_r is the current in the r^{th} terminal pair of \mathbf{N}_1 and ℓ_r that in the r^{th} pair of \mathbf{N}_2 , $1 \leq r \leq n$. Let \mathbf{K} be an n -dimensional space. Given an n -tuple $k \in \mathbf{K}$, say

$$k = [k_1, \dots, k_n],$$

we define the operator C from \mathbf{K} to \mathbf{J} by

$$j = Ck = [k_1, \dots, k_n, k_1, \dots, k_n].$$

Restricting $\bar{\mathbf{N}}$ by C gives the series combination \mathbf{N} of \mathbf{N}_1 and \mathbf{N}_2 . The details may easily be supplied by the interested reader.

2.41 Representing the series and parallel connections in terms of juxtaposition and restriction makes the lemma, 2.2, an immediate consequence of the lemma of (I, 17.2) and the theorems of (I, 17.3, 18.3).

2.5 We report here for record a curious property of PR operators which has so far found no application:

Lemma: If $Z(p)$ is a PR impedance operator from \mathbf{K} to $\mathbf{V} = \mathbf{K}^*$, and $Y(p)$ any PR admittance operator from \mathbf{V} to \mathbf{K} , then the operator

$$1 + Y(p)Z(p)$$

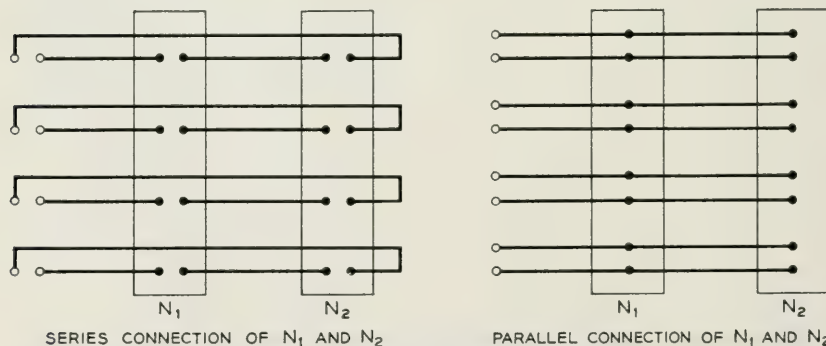


Fig. 2—Series and parallel connection of $2N$ poles: Series, left, and parallel, right.

in \mathbf{K} is non-singular. Dually

$$1 + Z(p)Y(p)$$

in \mathbf{V} is non-singular.

Proof: Suppose that $k \in \mathbf{K}$ is such that

$$(1 + Y(p)Z(p))k = 0 \quad (1)$$

for all p . Then

$$0 = Z(\bar{p})(1 + Y(p)Z(p))k = Z(\bar{p})k + Z(\bar{p})Y(p)Z(p)k$$

for all p . Then, however,

$$(Z(\bar{p})k, k) + (Z(\bar{p})Y(p)Z(p)k, k) = 0. \quad (2)$$

We may write the second term as

$$\overline{(Z^*(\bar{p})k, Y(p)Z(p)k)} = (Z(p)k, Y^*(p)Z^*(\bar{p})k) \quad (3)$$

by (I, 14.0) applied twice. Now $Z(p)$ is PR, in particular real and symmetric, so

$$Z^*(\bar{p}) = \overline{Z^*(p)} = Z'(p) = Z(p).$$

Using a similar calculation with $Y(p)$, the quantity (3) becomes

$$(Z(p)k, Y(\bar{p})Z(p)k). \quad (4)$$

For each $p \in \Gamma_+$, we have $\bar{p} \in \Gamma_+$ and the first term of (2) has a non-negative real part. But for $\bar{p} \in \Gamma_+$, (4) is the conjugate of

$$(v, Y(\bar{p})v) \quad (5)$$

where $v = Z(p)k$. Now (5) is a PR function of \bar{p} , hence has a non-negative real part for $\bar{p} \in \Gamma_+$, for any v . In particular therefore this is true for the v which, at p , makes (5) the conjugate of (4). Therefore (4) has a non-negative real part throughout Γ_+ . It follows from (2) then that

$$\operatorname{Re}(Z(\bar{p})k, k) = 0$$

for all $\bar{p} \in \Gamma_+$. By 2.02, then,

$$Z(p)k \equiv 0.$$

By (1), then

$$1k = k = 0.$$

Hence (1) implies $k = 0$. Therefore the operator in (1) has an inverse

III. A SIMPLE REALIZABILITY THEOREM

3.0 The following theorem is contained in Cauer⁵. Since it provides the basic step in our realizability process, we shall prove it here.

3.1 *Theorem:* Let $f(p)$ be any one of the four functions

$$(i) f(p) \equiv 1,$$

$$(ii) f(p) \equiv p,$$

$$(iii) f(p) = \frac{1}{p}$$

$$(iv) f(p) = \frac{2p}{p^2 + \omega_0^2}, \quad \omega_0^2 > 0.$$

Let R be a real, constant, symmetric semidefinite $n \times n$ matrix of rank r . Then:

(A) The matrix

$$Z(p) = f(p)R$$

is PR and there exists a finite passive $2n$ -pole \mathbf{N} with the impedance matrix $Z(p)$.

(B) The $2n$ -pole \mathbf{N} can be realized with ideal transformers and, respectively,

- (i) with r resistors,
- (ii) with r coils,
- (iii) with r capacitors,
- (iv) with r coils and r capacitors.

(C) The dual statements to (A) and (B) are true.

Proof: That $Z(p)$ in PR is easily verified directly. It will follow also from the results of Part I when we exhibit a (finite passive) network whose matrix is $Z(p)$. To construct this latter, let D be a diagonal matrix such that

$$R = WDW'$$

where W is a real, constant, non-singular matrix. That D and W always exist is the analog for impedance operators of the result of Halmos⁹, par. 41, for dimensionless operators. In fact, W can be taken to be orthogonal ($W^{-1} = W'$, cf. Halmos⁹, par. 63). If R is of rank r , D has r non-vanishing diagonal elements, say d_{11} , d_{22} , \dots , d_{rr} .

Since R is semidefinite, each $d_{ii}f(p)$, $1 \leq i \leq r$, is the impedance of an obviously passive two pole. Call this two-pole \mathbf{M}_i . Let $\mathbf{M}_{r+1}, \dots, \mathbf{M}_n$ be two poles consisting of short circuits. Consider the $2n$ -pole \mathbf{N}_1

made by connecting \mathbf{M}_s between T_s and T'_s , $1 \leq s \leq n$. This $2n$ -pole has the impedance matrix

$$Z_1(p) = f(p)D.$$

Then

$$Z(p) = f(p)WDW' = WZ_1(p)W'$$

is the matrix of a $2n$ -pole \mathbf{N} which can be obtained from \mathbf{N}_1 by the use of ideal transformers. Clearly \mathbf{N}_1 , and therefore \mathbf{N} , contains exactly the elements claimed in (B) of the theorem.

The dual theorem (C) needs no comment.

3.11 Corollary: The conclusion (A) of 3.1 holds if the hypotheses on $f(p)$ are replaced by " $f(p)$ is PR." The same method of proof applies but one must use the Brune theory to realize the impedances $d_{ii}f(p)$, $1 \leq i \leq r$.

3.2 The case (ii) of 3.1 shows that any physical system of coupled coils can be realized with a set of isolated (i.e., not coupled) coils, with ideal transformers to supply the coupling [Cf. (I, 19.12)]. With this fact in mind, we see that the method of network synthesis used in (I, 19) can be simplified to the following: one starts with a finite collection of two-poles: each one is a resistor, capacitor, or coil (inductor). These are then appropriately connected to suitable ideal transformers. Viewed from certain selected terminals of these transformers, this network is a $2n$ -pole equivalent to the desired one.

The difference between this process and that of (I, 19) is the minor one that coupled coils have been eliminated. We may then, however, regard any finite passive network as made up solely of simple two-poles (resistors, capacitors, coils) and ideal transformers.

It is readily verified from (I, 19.2) that open and short circuits are special cases of ideal transformers.

If a network made up in this way uses ℓ coils and c capacitors, we shall call $\ell + c$ the number of reactive elements in the network (or used by, or used in, the network).*

3.21 Lemma: The network described in the proof of 3.1 uses $\delta(Z)$ reactive elements. This is obvious from 2.12, 2.15, and 2.16.

IV. THE BRUNE PROCESS FOR A POSITIVE REAL MATRIX

4.0 Let $Z(p)$ be an $n \times n$ PR matrix. We can regard it as the impedance matrix of a general $2n$ -pole \mathbf{N} . In this section we shall describe the

* By this definition, a reactive element is an energy storage element. Ideal transformers are not reactive, by the very fact of their ideality.

construction of a finite passive network which, as a $2n$ -pole, has the impedance matrix $Z(p)$ —i.e. is a $2n$ -pole equivalent to \mathbf{N} . We call such a network a (physical) realization of \mathbf{N} , or of $Z(p)$. The dual problem, that of realizing a PR admittance matrix, can be handled dually.

Let $Z_0(p) = Z(p)$, $\mathbf{N}_0 = \mathbf{N}$, $n_0 = n$. We describe an inductive procedure which, given a $2n_r$ -pole \mathbf{N}_r , $r \geq 0$, either

- (i) Constructs a physical realization of \mathbf{N}_r , or
- (ii) Constructs a $2n_{r+1}$ -pole \mathbf{N}_{r+1} such that if \mathbf{N}_{r+1} is physically realizable, then \mathbf{N}_r is.

To show that this induction actually gives a realization of any PR matrix $Z_0(p)$ we must demonstrate that, first, it is effective—i.e. that at any stage \mathbf{N}_r at least one of (i) and (ii) is possible. Second, we must show that the process terminates with the construction of a finite network. The details of these demonstrations are given in the paragraphs 4.1 et seq. of this section. In the paragraphs 4.01 to 4.07 we describe the logical pattern into which these details are to be fit when they are established.

4.01 There are nine basic operations by which the networks \mathbf{N}_r are constructed. We name the operations here, in order to give a clearer picture of the logic of the process, but their mathematical treatment is deferred to later paragraphs.

IP: A PR impedance matrix $Z_r(p)$ which has poles on $p = i\omega$ is represented as

$$Z_r(p) = pR_\infty + \frac{1}{p}R_0 + \sum_k \frac{2p}{p^2 + \omega_k^2} R_k + Z_{r+1}(p),$$

where $Z_{r+1}(p)$ is PR and has no poles on $p = i\omega$.

AP: A PR admittance matrix $Y_r(p)$ is represented dually:

$$Y_r(p) = pG_\infty + \frac{1}{p}G_0 + \sum_k \frac{2p}{p^2 + \omega_k^2} G_k + Y_{r+1}(p).$$

ID: A PR impedance matrix $Z_r(p)$ is represented as $W'Z_{r+1}^B(p)W$, where $Z_{r+1}^B(p)$ is a non-singular $Z_{r+1}(p)$ bordered by zeros.

AD: Dual to ID.

Res: A PR matrix $Z_r(p)$ is represented as

$$Z_r(p) = aS + Z_{r+1}(p),$$

where S is real, constant, symmetric, and positive definite, and $a \geq 0$ is the largest a for which $Z_{r+1}(p)$ is PR.

Con: The dual to Res.

IB: This is the analog of the step in the Brune process for scalars in which the reactance of a minimum resistance structure is tuned out to create a zero. The details are intricate in the generalization to $2n$ -poles.

AB: This is the dual operation to IB.

F: A $2n_r$ -pole \mathbf{N}_r which has a constant PR matrix (admittance or impedance) is realizable at once, by 3.1. The operation F denotes this realization.

To each \mathbf{N}_r , one of these nine operations is to be applied. The effect of the last (F) is clearly salutary. That of each of the others is to split off a realizable piece of \mathbf{N}_r and leave a $2n_{r+1}$ -pole \mathbf{N}_{r+1} to which again some one of the operations is to be applicable.

Exactly which of these operations to apply at any stage depends upon the properties of the \mathbf{N}_r in question. We shall first devise a notation for describing the relevant properties of \mathbf{N}_r , and then in 4.04 present a table which summarizes what is to be proved in the paragraphs 4.1 et seq.

4.02 Definition: We say that $Z(p)$ has a zero of its real part at $p = i\omega_0$ if for some $k \in \mathbf{K}$, $k \neq 0$, we have

$$[Z(i\omega_0) + Z(-i\omega_0)]k = 0.$$

4.03 Let I be an integer describing a $2n$ -pole \mathbf{N} as follows:

$I = 0$ if \mathbf{N} has no impedance matrix.

$I = 1$ if \mathbf{N} has a non-constant impedance matrix which has no poles on $p = i\omega$, and no zeros of its real part on $p = i\omega$.

$I = 2$ if \mathbf{N} has a non-constant impedance matrix with a zero of its real part on $p = i\omega$, but no poles on $p = i\omega$.

$I = 3$ if \mathbf{N} has an impedance matrix with a pole or poles on $p = i\omega$.

Let A be an integer describing the admittance category of \mathbf{N} in a dual way (e.g., $A = 0$ if \mathbf{N} has no admittance matrix, etc.).

Let (I, A) denote the category of $2n$ -poles \mathbf{N} for which the indicated values of both I and A are true. Let

$$(I_1 + I_2, \quad A_1 + A_2) \tag{1}$$

denote the category of $2n$ -poles \mathbf{N} for which either I_1 or I_2 is true and, simultaneously, either A_1 or A_2 is true, with a similar definition for more summands. Then for example the category (1) above consists of the logical union of the following:

$$(I_1, A_1), \quad (I_1, A_2), \quad (I_2, A_1), \quad (I_2, A_2).$$

Let C denote the category of $2n$ -poles \mathbf{N} which have a constant matrix, impedance or admittance.

It is clear that any $2n$ -pole \mathbf{N} belongs in C or in exactly one of the sixteen elementary categories whose union is $(0 + 1 + 2 + 3, 0 + 1 + 2 + 3)$.

Table 4.04 shows for each category of \mathbf{N}_r , except $(0, 0)$, which operations may be applied, and the possible categories of the resulting \mathbf{N}_{r+1} .

A $2n$ -pole \mathbf{N} not in $(0, 0)$ has at least one matrix, and if it has two these are of the same degree (2.07, 2.13). We may then denote the degree of whatever matrix \mathbf{N} has simply by $\delta(\mathbf{N})$. The fourth and fifth columns of Table 4.04 show the relations of $\delta(\mathbf{N}_r)$ to $\delta(\mathbf{N}_{r+1})$, and of n_r to n_{r+1} .

4.05 Table 4.04 summarizes facts to be proved in 4.1 et seq. Assuming now that the assertions in this table are true, we can show that the inductive procedure is effective.

We observe first that the category C and every possible elementary category (I, A) except $(0, 0)$ is contained in at least one of the categories listed in the first column. Hence to any $2n$ -pole not in $(0, 0)$ there is at least one operation applicable. Further we note that the category $(0, 0)$ does not appear in the third column. Since by hypothesis \mathbf{N}_0 is not in $(0, 0)$, it follows by induction that no \mathbf{N}_r will be. Therefore the process can stop only by the operation F: completion.

Second, we notice that if \mathbf{N}_r is not in the category $(1, 1)$, then an applicable operation can be found which actually reduces one of the two numbers $\delta(\mathbf{N}_r)$, n_r . Furthermore, if \mathbf{N}_r is in $(1, 1)$, a sequence of two operations can be found which reduces one of $\delta(\mathbf{N}_r)$, n_r . Therefore before the realization process terminates (with F),

- (i) There are not more operations chosen from the list IP, AP, IB, AB, than the integer $\delta(\mathbf{N}_0)$;
- (ii) There are not more operations chosen from the list ID, AD, than the integer $n_0 - 1$ (since after these, still $n_{r+1} > 0$);

TABLE 4.04

Category of \mathbf{N}_r	Operation	Category of \mathbf{N}_{r+1}	$\delta(\mathbf{N}_r) - \delta(\mathbf{N}_{r+1})$	$n_r - n_{r+1}$
$(3, 0 + 1 + 2 + 3)$	IP	$C + (1 + 2, 0 + 1 + 2 + 3)$	> 0	0
$(0 + 1 + 2 + 3, 3)$	AP	$C + (0 + 1 + 2 + 3, 1 + 2)$	> 0	0
$(1 + 2, 0)$	ID	$(1 + 2, 1 + 2 + 3)$	0	$> 0^*$
$(0, 1 + 2)$	AD	$(1 + 2 + 3, 1 + 2)$	0	$> 0^*$
$(1, 1)$	Res	$(2, 0 + 1 + 2 + 3)$	0	0
$(1, 1)$	Con	$(0 + 1 + 2 + 3, 2)$	0	0
$(2, 1 + 2)$	IB	$C + (1 + 2 + 3, 0 + 1 + 2 + 3)$	> 0	0
$(1 + 2, 2)$	AB	$C + (0 + 1 + 2 + 3, 1 + 2 + 3)$	> 0	0
C	F	—	—	—

* But $n_{r+1} > 0$.

(iii) There are not more operations chosen from the list Res, Con, than the integer $\delta(\mathbf{N}_0) + n_0 - 1$.

Finally, then, the process must terminate after at most $2\delta(\mathbf{N}_0) + 2n_0 - 1$ operations.

4.06 Besides the data in 4.04, one other fact must be established about each operation: that \mathbf{N}_r is physically realizable if \mathbf{N}_{r+1} is. This will be done as we discuss each operation. When it is established, we reason back from the result of operation F, which provides a physical realization of some \mathbf{N}_m ($m \leq 2\delta(\mathbf{N}_0) + n_0 - 1$), through \mathbf{N}_{m-1} to $\mathbf{N}_0 = \mathbf{N}$, and obtain a realization of \mathbf{N} in finitely many steps.

4.07 Finally, we shall prove about each step that:

If \mathbf{N}_{r+1} can be realized with x_{r+1} reactive elements, then \mathbf{N}_r can be realized with

$$x_{r+1} + \delta(\mathbf{N}_r) - \delta(\mathbf{N}_{r+1})$$

reactive elements. This observation will provide the basis for proving the theorem of 2.18. For if \mathbf{N}_m is the network with which the process terminates, then by 3.21 \mathbf{N}_m can be realized with $\delta(\mathbf{N}_m)$ reactive elements. Reading back through the construction, each increment of degree that is encountered is balanced by an equal increment in the total number of reactive elements, so that, finally, $\delta(\mathbf{N})$ is the total number of reactive elements used. That no construction using fewer reactive elements can succeed will be shown in Section 6.

We now turn to IP, ID, Res, and IB, omitting the dual considerations. In each case, notation is simplified by writing Z , Y , \mathbf{N} , n respectively for Z_r , Y_r , \mathbf{N}_r , n_r , and Z_1 , Y_1 , \mathbf{N}_1 , n_1 for Z_{r+1} , Y_{r+1} , \mathbf{N}_{r+1} , n_{r+1} .

4.1 Given a $2n$ -pole \mathbf{N} in any category for which $I = 3$, its impedance matrix $Z(p)$ exists by hypothesis and has poles on $p = i\omega$. These can be removed successively by 2.05 and 2.06, giving

$$Z(p) = pR_\infty + \frac{1}{p}R_0 + \sum_{k=1}^K \frac{2p}{p^2 + \omega_k^2} R_k + Z_1(p). \quad (1)$$

In this expansion, either of R_0 , R_∞ may of course be absent, and all the R_k are real, symmetric, constant and semidefinite, for $k = 0, 1, \dots, K, \infty$. Furthermore, $Z_1(p)$ is PR and has no poles on $p = i\omega$, by 2.05, 2.06 and construction.

Let \mathbf{N}_1 be the $2n_1$ -pole whose impedance matrix is $Z_1(p)$. We define IP to be the operation giving \mathbf{N}_1 from \mathbf{N} . Either $\mathbf{N}_1 \in C$, or $I = 1$ or 2

for \mathbf{N}_1 , since at least $Z_1(p)$ exists. Furthermore, by construction $Z_1(p)$ is again an $n \times n$ matrix, so $n_1 = n$.

By 2.14 and 2.15,

$$\delta(Z) = \text{rank}(R_\infty) + \text{rank}(R_0) + 2 \sum_{k=1}^K \text{rank}(R_k) + \delta(Z_1). \quad (2)$$

Since $\delta(Z)$ is finite, this shows that K is finite. Indeed, $2K \leq \delta(Z)$. Furthermore, $\delta(Z) > \delta(Z_1)$, because a matrix of rank zero is itself zero, and by hypothesis $Z(p)$ has a pole on $p = i\omega$. Therefore we have established the claims in the first line of the Table 4.04, and by a dual argument those in the second line.

We must yet show that if \mathbf{N}_1 is physically realizable, then \mathbf{N} is. Each term in (1), save $Z_1(p)$, is the matrix of a physically realizable $2n$ -pole, by 3.1. There are at most $K + 2$ such terms. The series combination of their respective $2n$ -poles is therefore physically realizable and \mathbf{N} results from the series connection of these and \mathbf{N}_1 (2.2). Therefore if \mathbf{N}_1 is realizable, so is \mathbf{N} .

Fig. 3 shows the relation of \mathbf{N} and \mathbf{N}_1 under IP, and the dual relation under AP. Here we have shown $n = 3$. The boxes labelled 0, ∞ , \dots , K are the devices corresponding to the poles at 0, ∞ , \dots , $i\omega_K$, the terminals on the extreme left are those of \mathbf{N} , and \mathbf{N}_1 is on the right.

4.11 From (2), and (B) of 3.1, we see that the number of reactive elements used in the realization of the network between \mathbf{N}_1 and \mathbf{N} is exactly

$$\delta(Z) - \delta(Z_1) = \delta(\mathbf{N}) - \delta(\mathbf{N}_1).$$

Clearly the dual result holds for AP. This verifies 4.07 for IP and AP.

4.2 Consider a $2n$ -pole \mathbf{N} in $(1 + 2, 0)$. In particular, then, the impedance matrix $Z(p)$ of \mathbf{N} exists and is not constant, but $Z(p)$ has no inverse. Then 2.08 applies, and we have

$$Z(p) = W'Z_1^B(p)W, \quad (1)$$

where W is real, constant, and non-singular, and $Z_1^B(p)$ is a non-singular matrix $Z_1(p)$ bordered by zeros. Let \mathbf{N}_1 be the $2n_1$ -pole whose impedance matrix is $Z_1(p)$. We define ID as the operation which gives \mathbf{N}_1 from \mathbf{N} . Now $n_1 < n$, because $Z(p)$ is singular and $Z_1(p)$ is not. Also, $Z_1(p)$ is not constant, because $Z(p)$ is not, and $\delta(Z_1) = \delta(Z)$, by 2.17. Therefore $n_1 \neq 0$, also \mathbf{N}_1 is not in C . Because $Z_1(p)^{-1}$ exists, \mathbf{N}_1 is in $A = 1, 2$ or 3. Because $Z(p)$ has no poles on $p = i\omega$, neither has $Z_1(p)$, so $\mathbf{N}_1 \in (1 +$

2, 1 + 2 + 3). This verifies the statements on the third line of the Table 4.04, and the fourth by duality.

That \mathbf{N} is physically realizable if \mathbf{N}_1 is, is the gist of (I, 8.11) and (I, 8.4). We prove it here by noting that $Z_1^H(p)$ is the matrix of a $2n$ -pole \mathbf{N}_2 which obtains by adjoining $n - n_1 > 0$ pairs of shorted terminals to \mathbf{N}_1 . Then (1) shows that \mathbf{N} obtains from \mathbf{N}_2 by the use of ideal transformers (I, 9.1).

Fig. 4 shows in schematic form the effects of the operation ID and AD. In each case, it is emphasized that \mathbf{N}_1 has a matrix dual to that of \mathbf{N} . We have shown $n = 5$, $n_1 = 3$.

4.21 No reactive elements are used in this construction, so 4.07 is satisfied.

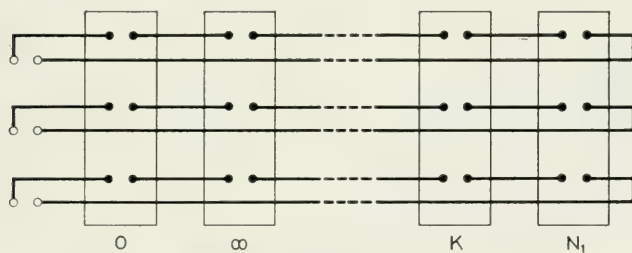
4.3 Consider now a $2n$ -pole \mathbf{N} in (1, 1). Then its impedance matrix $Z(p)$ is finite for every $p = i\omega$, and not constant.

Let $R(p)$, $I(p)$, respectively, be the real and imaginary parts of $Z(p)$:

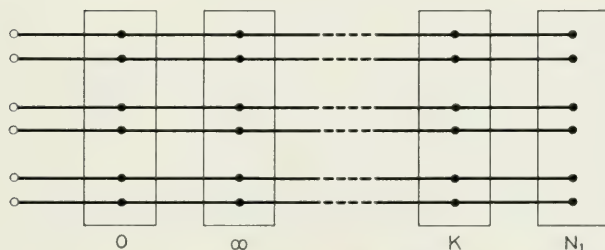
$$2R(p) = Z(p) + \overline{Z(\bar{p})} = Z(p) + Z(\bar{p}) = Z(p) + Z^*(p);$$

$$2iI(p) = Z(p) - \overline{Z(\bar{p})} = Z(p) - Z(\bar{p}) = Z(p) - Z^*(p);$$

$$Z(p) = R(p) + iI(p).$$



STRUCTURE RESULTING FROM IP



STRUCTURE RESULTING FROM AP

Fig. 3—Structure resulting from IP, above and AP, below.

Then $R(p) = R^*(p)$, $I(p) = I^*(p)$, and both are real and symmetric. If k is any vector,

$$(Z(p)k, k) = (R(p)k, k) + i(I(p)k, k),$$

and the self-adjoint property of R and I imply that each scalar product on the right is real. Therefore

$$\begin{aligned} \operatorname{Re}(Z(p)k, k) &= (R(p)k, k), \\ \operatorname{Im}(Z(p)k, k) &= (I(p)k, k). \end{aligned} \quad (1)$$

We note that

$$2iI(\bar{p}) = Z(\bar{p}) - Z^*(\bar{p}) = Z^*(p) - Z(p) = -2iI(p)$$

so that, in particular, $I(i\omega)$ is an odd function of ω .

4.31 *Lemma:* Let S be a given real, constant, symmetric, and positive definite matrix. Then there exists a unique number $a > 0$ such that

(i) The matrix

$$R(i\omega) - aS$$

is semidefinite for every ω ,

(ii) For some $\omega_0 \geq 0$, possibly $+\infty$,

$$R(i\omega_0) - aS$$

is singular.

Proof: We first show how the number a would be calculated, and then reduce the claims of the lemma to a well-known and basic theorem in

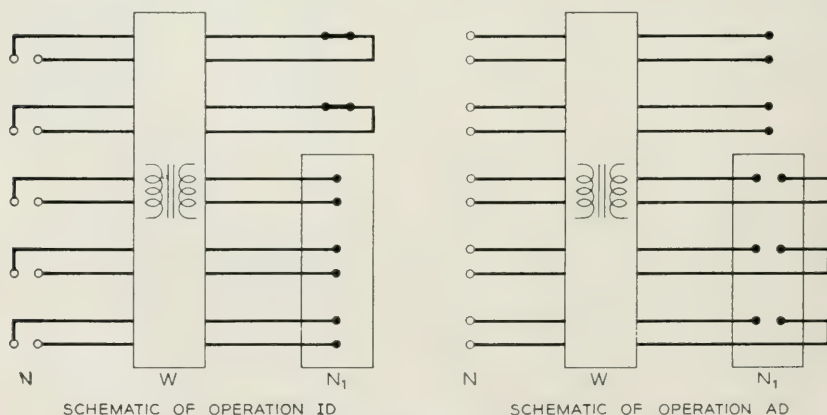


Fig. 4—Schematic of operation ID, left and AD, right.

the theory of quadratic forms. Fix ω and consider the matrix

$$R(i\omega) - \lambda S$$

as a function of λ . Its determinant,

$$\Delta_\omega(\lambda) = |R(i\omega) - \lambda S|,$$

is an n^{th} degree polynomial in λ with the following two properties:

- (α) The coefficient of λ^n in $\Delta_\omega(\lambda)$ is not zero and is independent of ω ,
- (β) The n roots of

$$\Delta_\omega(\lambda) = 0 \tag{2}$$

are real and positive.

Now $R(i\omega)$ is rational, hence continuous, and finite for all ω , including $\omega = \infty$, by the hypothesis that \mathbf{N} is in $(1, 1)$. By (α) above, therefore, each root of (2) is a continuous function of ω on the compact set $-\infty \leq \omega \leq \infty$. Let $a(\omega)$ denote the least root of (2). Then $a(\omega)$ is again bounded and continuous for all ω . There is, therefore, an ω_0 where $a(\omega)$ takes its least value. This is the ω_0 referred to in the lemma, and

$$a = a(\omega_0).$$

We see that this calculation requires solving an n^{th} degree polynomial equation containing a parameter (ω), and then minimizing the least root by varying the parameter. Though some properties of $R(i\omega)$ are available to assist in the process, and the choice of S is somewhat free to us, this is scarcely a feasible calculation in practice. Even when one reduces the minimizing problem to finding the roots of a derivative, there remains a prodigious calculation in all but the simplest cases.

Since by its definition $R(i\omega) = R(-i\omega)$, we may take $\omega_0 \geq 0$.

The relation (1) above implies that

$$(R(i\omega)k, k) \geq 0$$

for all real ω and all $k \in \mathbf{K}$, because $Z(p)$ is PR. That is, $R(i\omega)$ is semi-definite. The hypothesis that $Z(p)$ has no zero of its real part $R(p)$ on $p = i\omega$ then implies that $R(i\omega)$ is positive definite. All of (i), (ii), (α), and (β) then follow from well-known properties of definite quadratic forms. They may, for example, all be deduced from Halmos⁹, paragraphs 62, 63, and 74, by choosing a coordinate frame in which the operator corresponding to S above is represented by the unit matrix. A more elegant reduction to the cited results of Halmos⁹ can also be constructed.

4.32 Lemma: Given \mathbf{N} in $(1, 1)$, we choose any real constant symmetric

and positive definite matrix S and find the a described in 4.31. Then the matrix

$$Z_1(p) = Z(p) - aS$$

is PR and has a zero of its real part at $p = i\omega_0$.

Proof: Clearly $Z_1(p)$ is symmetric. By 2.09, then, $Z_1(p)$ is PR if the function

$$\varphi_1(p) = (Z_1(p)k, k) = (Z(p)k, k) - a(Sk, k) \quad (3)$$

is PR for each k . Clearly this function is rational and has no singularities in Γ_+ . It suffices then to show that its real part is non-negative on $p = i\omega$. By (1) of 4.3

$$\operatorname{Re} \varphi_1(i\omega) = (R(i\omega)k, k) - a(Sk, k)$$

and this is non-negative by (i) of 4.31.

That $Z_1(p)$ has a zero of its real part at $p = i\omega_0$ is (ii) of 4.31.

4.33 Let \mathbf{N}_1 be the $2n$ -pole whose impedance matrix is the $Z_1(p)$ of 4.32. We define the operation Res as that which produces \mathbf{N}_1 from \mathbf{N} . It is evident from (3) above that the poles of $Z_1(p)$ are exactly those of $Z(p)$, hence $I = 2$ for \mathbf{N}_1 . Nothing can be said of the admittance matrix for \mathbf{N}_1 . $\delta(Z_1) = \delta(Z)$ by 2.14 and 2.15, and $n_1 = n$ by construction. The claims in 4.04 are now established for Res, and dually for Con.

The relation

$$Z(p) = Z_1(p) + aS$$

shows that \mathbf{N} is a series combination of \mathbf{N}_1 and a device with the impedance matrix aS . Since $a > 0$, this latter is a realizable resistance network (3.1). Hence \mathbf{N} is realizable if \mathbf{N}_1 is.

4.34 We observe that no reactive elements are used in the network between \mathbf{N}_1 and \mathbf{N} (2.12, 3.12). This verifies 4.07 for Res and Con.

4.4 We now turn to the *piece de resistance* of the generalized Brune process, the operation IB and its dual. Consider a $2n$ -pole \mathbf{N} in the category $(2, 1 + 2)$ —i.e., its impedance matrix $Z(p)$ exists, is not constant, is non-singular on $p = i\omega$, and has a zero of its real part at some $p = i\omega_0$. We have for some $k \in \mathbf{K}$ such that $k \neq 0$,

$$R(i\omega_0)k = 0. \quad (1)$$

Here, $R(p)$ is as defined in 4.3.

4.41 We now assert that we may assume that $0 < \omega_0$, and $i\omega_0 \neq \infty$ in

(1). Certainly we may take $\omega_0 \geq 0$, because $R(i\omega) = R(-i\omega)$. Furthermore, by (1),

$$Z(i\omega_0)k = iI(i\omega_0)k. \quad (2)$$

$I(i\omega)$, being odd, and finite everywhere on $p = i\omega$, must vanish at $\omega = 0$, and at $i\omega = \infty$. Hence if $\omega_0 = 0$ or $i\omega_0 = \infty$, $Z(i\omega_0)k = 0$ and $Z(p)$ is singular on $p = i\omega$. This denies our hypothesis that $\mathbf{N} \in (2, 1 + 2)$.

4.42 Let \mathbf{J} be the set of all vectors $k \in \mathbf{K}$ such that (1) holds: the null space of $R(i\omega_0)$. Then clearly \mathbf{J} is a linear manifold. Furthermore, \mathbf{J} is real, because, if (1) holds then

$$\overline{R(i\omega_0)k} = \overline{R(i\omega_0)\bar{k}} = R(i\omega_0)\bar{k} = \bar{0} = 0$$

and \bar{k} also is in \mathbf{J} .

Relations (1) and (2) hold for all $k \in \mathbf{J}$.

4.43 By its construction, $I(i\omega_0)$ is real and symmetric, but not necessarily definite. There does however exist a real diagonal matrix D and a real non-singular W such that $I(i\omega_0) = W'DW$. Let D_+ be the (diagonal) matrix obtained from D by replacing all negative elements of D by zero, and define D_- by

$$D = D_+ - D_-. \quad (3)$$

Then D_+ and D_- are real, symmetric, and non-negative semidefinite. Define

$$\begin{aligned} A &= \omega_0 W'D_+W, \\ B &= \frac{1}{\omega_0} W'D_-W. \end{aligned} \quad (4)$$

We have chosen $\omega_0 > 0$, so A and B are both real, symmetric and non-negative. Certainly therefore

$$Z^{(2)}(p) = Z(p) + \frac{1}{p}A + pB \quad (5)$$

is PR. Also $Z^{(2)}(p)$ has an inverse, because $Z(p)$ has one by hypothesis and 2.2 applies.

4.431 Let $v \in \mathbf{V}$ be such that for some $k_1 \in \mathbf{K}$

$$v = Ak_1$$

and for some $k_2 \in \mathbf{K}$

$$v = Bk_2.$$

Then $v = 0$.

Proof: We may assume that the first r diagonal elements of D are the non-zero elements of D_+ , the next s those of $-D_-$. By (4),

$$(W')^{-1}v = \omega_0 D_+ W k_1,$$

$$(W')^{-1}v = \frac{1}{\omega_0} D_- W k_2.$$

The first of these relations exhibits $(W')^{-1}v$ as an n -tuple with non-zero components at most among the first r , the second as an n -tuple with non-zero components at most among the last $n - r$. Hence all components of $(W')^{-1}v$ are zero. Hence v itself is zero.

4.44 Define

$$X(p) = -\frac{1}{p} A - pB, \quad (6)$$

and let \mathbf{N}_X be the $2n$ -pole whose impedance matrix is $X(p)$. \mathbf{N}_X is not physically realizable, since it is made up of negative reactances.

Let $\mathbf{N}^{(2)}$ be the $2n$ -pole whose impedance matrix is $Z^{(2)}(p)$. Then by (5) \mathbf{N} obtains from $\mathbf{N}^{(2)}$ and \mathbf{N}_X by connecting them in series.

We have the following relation holding on $p = i\omega$, but only thereon since it is only there that $X(p)$ is a pure imaginary:

$$Z^{(2)}(i\omega) = R(i\omega) + i \left[I(i\omega) - \frac{1}{\omega} A + \omega B \right].$$

In particular, at $i\omega_0$,

$$\begin{aligned} Z^{(2)}(i\omega_0) &= R(i\omega_0) + i[I(i\omega_0) - W'D_+W + W'D_-W] \\ &= R(i\omega_0), \end{aligned}$$

by (3) and (4). Since \mathbf{J} is the null space of $R(i\omega_0)$ by definition, \mathbf{J} is the null space of $Z^{(2)}(i\omega_0)$.

4.45 Now $Y^{(2)}(p) = [Z^{(2)}(p)]^{-1}$ exists and is PR. Since $Z^{(2)}(i\omega_0)$ annihilates every element of \mathbf{J} , it follows that $Y^{(2)}(p)$ does not exist at $p = i\omega_0$ —therefore $Y^{(2)}(p)$ has a pole at $i\omega_0$. Hence we may apply AP and represent $Y^{(2)}(p)$ as a reactance network, with admittance matrix

$$G(p) = \frac{2p}{p^2 + \omega_0^2} G, \quad (7)$$

in parallel with a $2n$ -pole $\mathbf{N}^{(3)}$ which has an admittance matrix, say

$$Y^{(2)}(p) = G(p) + Y^{(3)}(p), \quad (8)$$

where $Y^{(3)}(p)$ is finite at $p = i\omega_0$.

4.46 Multiplying (8) on either side by $Z^{(2)}(p)$,

$$\begin{aligned} \frac{2p}{p^2 + \omega_0^2} GZ^{(2)}(p) + Y^{(3)}(p)Z^{(2)}(p) &= 1 \\ &= \frac{2p}{p^2 + \omega_0^2} Z^{(2)}(p)G + Z^{(2)}(p)Y^{(3)}(p). \end{aligned} \quad (9)$$

Here, to be strictly correct, we should write two separate equations, interpreting 1 as the identity operator in \mathbf{K} for, here, the left equality, and as the identity operator in \mathbf{V} for the right equality. Multiplying (9) through by $p - i\omega_0$ and letting $p \rightarrow i\omega_0$, we obtain

$$GZ^{(2)}(i\omega_0) = 0 = Z^{(2)}(i\omega_0)G.$$

Here, as in (9), we have condensed two dimensionally incompatible equalities. From this it follows that each of G and $Z^{(2)}(i\omega_0)$ has its range in the null space of the other. In particular, therefore, the range of G is contained in \mathbf{J} .

4.47 Consider now a v such that $Gv = 0$. Then, by (7) and (8),

$$v \equiv Z^{(2)}(p)Y^{(2)}(p)v \equiv Z^{(2)}(p)Y^{(3)}(p)v$$

so, at $i\omega_0$,

$$v = Z^{(2)}(i\omega_0)Y^{(3)}(i\omega_0)v = Z^{(2)}(i\omega_0)k$$

for some finite vector $k = Y^{(3)}(i\omega_0)v$. Since $Z^{(2)}(i\omega_0)$ is finite, $v \neq 0$ implies that $k \neq 0$. Then, however, v lies in the range of $Z^{(2)}(i\omega_0)$. Combining this fact with the result of 4.46, we see that for $Gv = 0$ it is necessary and sufficient that v lie in the range of $Z^{(2)}(i\omega_0)$: the range of $Z^{(2)}(i\omega_0)$ is exactly the null space of G .

4.48 Now in Halmos⁹, par. 37, it is shown that for any dimensionless operator in an n -space the dimensionality of its range space (its *rank*) and the dimensionality of its null space (its *nullity*) add up to n . A similar result and proof hold for operators between \mathbf{V} and \mathbf{K} . Let m be the dimensionality of \mathbf{J} . Then $n - m$ is the rank of $Z^{(2)}(i\omega_0)$, and therefore the dimensionality of the range of $Z^{(2)}(i\omega_0)$, and by 4.47 the dimensionality of the null space of G . Hence, finally,

$$\text{rank } (G) = n - (n - m) = m.$$

By 4.46, therefore, \mathbf{J} is exactly the range of G .

4.49 Now $\mathbf{N}^{(3)}$, whose admittance matrix is $Y^{(3)}(p)$, might not be ex-

pected to have an impedance matrix. The following reasoning shows that it does have, however:

Consider a $v \in \mathbf{V}$ for which $Y^{(3)}(p)v \equiv 0$. Then from the right side of (9), with (5),

$$v = \frac{2p}{p^2 + \omega_0^2} Z(p)Gv + \frac{2}{p^2 + \omega_0^2} AGv + \frac{2p^2}{p^2 + \omega_0^2} BGv. \quad (10)$$

We have by hypothesis that $Z(p)$ is finite on $p = i\omega$. Therefore we may calculate, by letting $p \rightarrow 0$ in (10), that

$$v = \frac{2}{\omega_0^2} AGv,$$

and, by letting $p \rightarrow \infty$ in (10), that

$$v = 2BGv.$$

These two equations exhibit v as an element in the range of A and also an element in the range of B . The only possible such v is $v = 0$, by 4.43. Therefore there is no non-zero v such that $Y^{(3)}(p)v \equiv 0$. Then $Z^{(3)}(p) = Y^{(3)}(p)^{-1}$ exists as a PR operator.

4.491 Let

$$L(p) = \frac{1}{p} H + pF \quad (11)$$

be the matrix whose poles at $p = 0$ and $p = \infty$ are those of $Z^{(3)}(p)$. That is, let

$$Z^{(3)}(p) = L(p) + Z^{(4)}(p), \quad (12)$$

where $Z^{(4)}(p)$ is PR and finite at 0 and ∞ . Because $Z^{(3)}(p)$ is PR, H and F are both real, symmetric, and semidefinite. Let \mathbf{N}_L be the $2n$ -pole whose impedance matrix is $L(p)$, and $\mathbf{N}^{(4)}$ the $2n$ -pole with matrix $Z^{(4)}(p)$. In fact, \mathbf{N}_L is realizable. $\mathbf{N}^{(3)}$ is the series combination of \mathbf{N}_L and $\mathbf{N}^{(4)}$, by (12).

4.5 Equations (5), (7), (8), and (12) above are statements about matrices in a particular coordinate frame—that frame appropriate to the given \mathbf{N} . We can interpret them as operator relations by simple decree. We wish now to draw a circuit diagram illustrating these relations. To do so, we introduce a suitable new coordinate frame.

Because $G(p)$ is PR and of rank m , we know that a frame can be found in which the matrix for $G(p)$ is an $m \times m$ non-singular matrix bordered by zeros (2.08, or (I, 16.8)). By (7) and the result of 4.48, we

may take the first m current vectors, k_1, k_2, \dots, k_m , specifying this frame, to span \mathbf{J} . It follows from the matrix form of G then that the corresponding dual vectors v_1, \dots, v_m span the range of G —i.e., the null space of $Z^{(2)}(i\omega_0)$. We shall adopt such a frame for the further discussion.

Let \mathbf{K}_1 be the space spanned by k_{m+1}, \dots, k_n , and \mathbf{V}_1 that spanned by v_{m+1}, \dots, v_n , in this frame. Then

$$\begin{aligned}\mathbf{K} &= \mathbf{J} \oplus \mathbf{K}_1 \\ \mathbf{V} &= \mathbf{U} \oplus \mathbf{V}_1,\end{aligned}\tag{1}$$

say, where $\mathbf{U} = \mathbf{J}^*$, $\mathbf{V}_1 = \mathbf{K}_1^*$ [Cf. (I, 10.6)].

If \mathbf{M} is the name of any given $2n$ -pole discussed in the paragraphs 4.4 to date, we let $\bar{\mathbf{M}}$ denote the Caue equivalent of \mathbf{M} in this new frame.

4.51 Let $\bar{\mathbf{N}}_G$ be the $2m$ -pole whose matrix in the new frame is the $m \times m$ non-singular admittance matrix which, when bordered, gives the matrix of the operator

$$G(p) = \frac{2p}{p^2 + \omega_0^2} G.$$

The $2n$ -pole whose matrix is $G(p)$ then obtains by adjoining $n-m$ open circuits to $\bar{\mathbf{N}}_G$. The matrix of $\bar{\mathbf{N}}_G$ operates from \mathbf{U} to \mathbf{J} and has an inverse.

4.52 Fig. 5 shows a diagram, which $n = 5$, $m = 3$, of the manner in which we now have \mathbf{N} represented. The terminals on the extreme left are those of \mathbf{N} . \mathbf{N} is obtained from $\bar{\mathbf{N}}$ by a transformer. The horizontal current paths cut the dotted section A-A at points which may be interpreted as the terminals of $\bar{\mathbf{N}}$. Ideal transformers, as in Fig. 1 of I, can be introduced here as needed. Putting them in the diagram merely complicates the picture.

$\bar{\mathbf{N}}$ is the series connection of $\bar{\mathbf{N}}_x$ and $\mathbf{N}^{(2)}$. The terminals of the latter are on B-B. $\mathbf{N}^{(2)}$, again, is the parallel connection of a $2n$ -pole obtained from $\bar{\mathbf{N}}_G$ by the adjunction of open circuits, and $\bar{\mathbf{N}}^{(3)}$. The latter has its terminals on C-C. Again, $\bar{\mathbf{N}}^{(3)}$ is the series connection of $\bar{\mathbf{N}}_L$ and $\bar{\mathbf{N}}^{(4)}$.

4.53 Let \mathbf{M}_{AD} be the device between A-A and D-D of Fig. 5. This device has n terminal pairs on A-A and n more on D-D. We may suppose that ideal transformers are attached at each terminal pair as in Fig. 1 of I, since including them in the construction of $\bar{\mathbf{N}}$ would not alter its behavior. Then \mathbf{M}_{AD} is a $2(2n)$ -pole.

\mathbf{M}_{AD} is constructed from certain $2r$ poles (with various r) as indicated in the diagram of Fig. 5. The ideal graph* of this diagram (rather, of

* Cf. (I, 4.1).

the relevant part of it between A-A and D-D) obtains from Fig. 5 by inserting ideal branches—two poles—across each terminal pair of each box, and neglecting the outlines of the boxes. The upper m channels of this ideal graph are then T sections, and the lower $n-m$ are degenerate T sections with no shunt arm. This ideal graph is shown in Fig. 6. The ideal branches are shown as small boxes.

The program of the next few paragraphs is to demonstrate that \mathbf{M}_{AD} is a physically realizable $2(2n)$ -pole.

4.54 Let us designate the terminal pairs of \mathbf{M}_{AD} on the section A-A by $T_1, T'_1, \dots; T_n, T'_n$, where the r^{th} pair is the intersection with A-A of the leads to the r^{th} terminal pair of $\bar{\mathbf{N}}$. We designate the pairs on

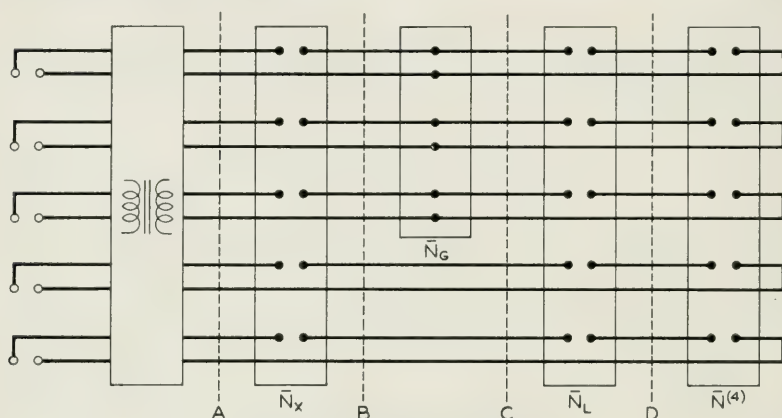


Fig. 5—Original form for $\bar{\mathbf{N}}$.

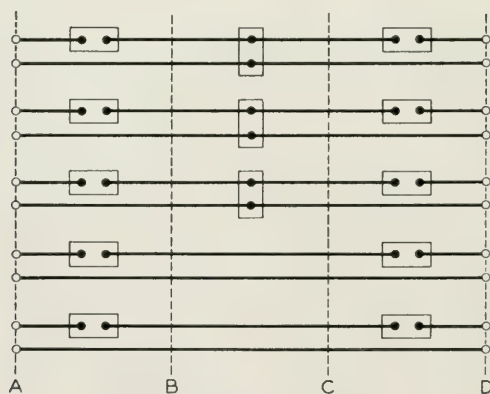


Fig. 6—Ideal graph of \mathbf{M}_{AD} .

D-D by $S_1, S'_1; \dots; S_n, S'_n$, where here the r^{th} pair is the intersection with D-D of the leads to the r^{th} terminal pair of $\bar{\mathbf{N}}^{(4)}$. In each case we orient the pair T, T' or S, S' so that the primed (negative) terminal is on the lead to the primed terminal of $\bar{\mathbf{N}}$ or $\bar{\mathbf{N}}^{(4)}$.

Let the $2n$ -tuple

$$[a_1, a_2, \dots, a_n, b_1, b_2, \dots, b_n] \quad (2)$$

represent the currents into the terminals of \mathbf{M}_{AD} in the order

$$T_1, T_2, \dots, T_n, S_1, \dots, S_n.$$

Then we may interpret

$$[a_1, \dots, a_n] \quad (3)$$

as a vector in \mathbf{K} expressed in the coordinate frame introduced for Fig. 5, and also

$$[b_1, \dots, b_n] \quad (4)$$

as a vector in \mathbf{K} in the same frame. That is, any current vector into \mathbf{M}_{AD} can be written as an ordered pair

$$k_1, k_2 \quad (5)$$

where each $k_i \in \mathbf{K}$, with the convention that such a pair determines a $2n$ -tuple (2) from the n -tuples (3) of k_1 and (4) of k_2 .

We shall write the ordered pair (5) in the form

$$k_1 \oplus k_2. \quad (6)$$

Because we have \mathbf{K} represented in the special way

$$\mathbf{K} = \mathbf{J} \oplus \mathbf{K}_1,$$

where \mathbf{J} is the subspace spanned by n -tuples (3) in which the last $n-m$ components vanish (this is (1) of 4.5) we can further split the $2n$ -tuple (2) into

$$(j_1 \oplus \ell_1) \oplus (j_2 \oplus \ell_2), \quad (7)$$

where $j_i \in \mathbf{J}$, $\ell_i \in \mathbf{K}_1$, $i = 1, 2$, and in (6)

$$k_i = j_i \oplus \ell_i. \quad (8)$$

Formulas dual to those of (2) through (8) of course hold for voltage $(2n)$ -tuples. Let \mathbf{K}^2 be the space of current $2n$ -tuples (2) (or (7)) and \mathbf{V}^2 the space of voltage $(2n)$ -tuples

$$[e_1, e_2, \dots, e_n, f_1, \dots, f_n] = (u_1 \oplus v_1) \oplus (u_2 \oplus v_2)$$

analogous to (2) and (7), with the scalar product

$$\sum_{r=1}^n e_r \bar{a}_r + \sum_{r=1}^n f_r \bar{b}_r. \quad (9)$$

It is a common and convenient malpractice in vector algebra to use, for example, the symbol j both for an m -tuple in \mathbf{J} and for the n -tuple

$$j \oplus 0 \in \mathbf{J} \oplus \mathbf{K}_1$$

of the form (8). Taking this advantage, we can see that (9) is simply

$$(u_1 + v_1, j_1 + \ell_1) + (u_2 + v_2, j_2 + \ell_2) \quad (9')$$

where here the parentheses denote scalar products between \mathbf{V} and \mathbf{K} . The form (9') can also be derived directly from (1), (7), and (I, 10.6).

4.55 We now wish to compute the voltage-current pairs admitted by \mathbf{M}_{AD} . Referring to Fig. 5, we observe that $\bar{\mathbf{N}}_X$ and $\bar{\mathbf{N}}_L$ both have impedance matrices ($X(p)$ and $L(p)$ respectively, or, rather, the matrix forms of these in the frame of present interest) finite at all p except $p = 0$, $p = \infty$. Each will, therefore, admit any current n -tuple into its terminals, i.e., through its ideal branches, at any but these exceptional frequencies. By construction, $\bar{\mathbf{N}}_G$ has a *non-singular* admittance matrix and therefore also will admit any current m -tuple into its terminals (2.07), except at most at certain isolated frequencies. It is evident by Kirchhoff's laws applied to Fig. 6 then that \mathbf{M}_{AD} will admit any current $2n$ -tuple of the form

$$(j_1 \oplus k) \oplus (j_2 \oplus (-k)) \quad (10)$$

where $j_i \in \mathbf{J}$, $i = 1, 2$, and $k \in \mathbf{K}_1$, except at most at finitely many exceptional frequencies. Conversely, if the current $2n$ -tuple specified by (7) is that in \mathbf{M}_{AD} , conservation at the absent shunt arms of the lower degenerate T-sections implies that, as elements of \mathbf{K} ,

$$k_1 + k_2 = 0,$$

that is, the current is of the form (10). Hence $2n$ -tuples of the form (10) span the space of currents admitted by \mathbf{M}_{AD} . Let us call this space \mathbf{K}_M^2 . It is a proper subspace of \mathbf{K}^2 unless $m = n$.

4.56 Let $G^{-1}(p)$ denote the $m \times m$ impedance matrix of $\bar{\mathbf{N}}_G$. Then by (7) of 4.4, interpreted as an operator equation,

$$G^{-1}(p) = \left(\frac{1}{2} p + \frac{\omega_0^2}{2p} \right) G^{-1} \quad (11)$$

where G^{-1} is a real, constant, symmetric, non-singular $m \times m$ matrix.

We can now compute the voltage across \mathbf{M}_{AD} corresponding to the current (10). Let w be the n -tuple of voltages appearing at the section B-B or C-C of Fig. 6, with its components listed in the appropriate order. Then we may interpret w as a vector in \mathbf{V} , and write it

$$w = u_0 \oplus v_0 \quad (12)$$

where $u_0 \in \mathbf{U}$, $v_0 \in \mathbf{V}_1$. Now by Kirchhoff's current law applied to the shunt arms in the upper channels of Fig. 6, the current into $\bar{\mathbf{N}}_a$ is

$$j_1 + j_2,$$

and therefore

$$u_0 = G^{-1}(p)(j_1 + j_2). \quad (13)$$

By Kirchhoff's voltage law applied to a typical mesh which begins on A-A, goes through \mathbf{N}_x to B-B, and then through a shunt arm and returns to A-A, the voltage n -tuple appearing at A-A is

$$X(p)(j_1 + k) + w.$$

Referring to (12), let us use u_0 also to denote the vector

$$u_0 \oplus 0 \in \mathbf{V},$$

and v_0 to denote

$$0 \oplus v_0 \in \mathbf{V}.$$

Interpreting (13) in this way we get

$$X(p)(j_1 + k) + G^{-1}(p)(j_1 + j_2) + v_0 \quad (14)$$

as the voltage n -tuple on A-A.

A similar calculation gives

$$L(p)(j_2 - k) + G^{-1}(p)(j_1 + j_2) + v_0 \quad (15)$$

as the voltage n -tuple on D-D. The ordered pair (14), (15) then gives the voltage $2n$ -tuple corresponding to (10), in the notation analogous to (5).

4.57 $X(p)$, $L(p)$, and $G^{-1}(p)$, respectively, are defined in (6) of 4.44, (11) of 4.491, and (11) of 4.56. Each one is finite except at $p = 0$ and $p = \infty$. Let $\Gamma_{\mathbf{M}}$ be the complex plane from which these two points are deleted. It is now possible to show that the linear correspondence whose pairs, for each $p \in \Gamma_{\mathbf{M}}$, are the voltages (14), (15) $\in \mathbf{V}^2$ and the currents (10) $\in \mathbf{K}^2$, satisfies P1 through P7 of (I, 6, 7)—that is, is PR (I, 16.71).

In the present special circumstances it is almost as easy to study \mathbf{M}_{AD} in a slightly different way than this. Since fewer direct references to \mathbf{I} are involved, we shall take the alternative path.

We first calculate the scalar product between the voltage (14), (15) and an arbitrary current of the form (10), say the current

$$(h_1 \oplus \ell) \oplus (h_2 \oplus (-\ell)) \in \mathbf{K}_{\mathbf{M}}^2.$$

To do so, we consider the form (9') for such a product. In the first writing, then, this scalar product is

$$\begin{aligned} (X(p)(j_1 + k) + G^{-1}(p)(j_1 + j_2) + v_0, h_1 + \ell) \\ + (L(p)(j_2 - k) + G^{-1}(p)(j_1 + j_2) + v_0, h_2 - \ell). \end{aligned}$$

Each of these scalar products has three voltages appearing in it. Distributing the products over these voltages, and using the facts that the range of $G^{-1}(p)$ is \mathbf{J} and that $v_0 \in \mathbf{V}_1 = (\mathbf{J})^0$ we get a second form:

$$\begin{aligned} (X(p)(j_1 + k), h_1 + \ell) + (G^{-1}(p)(j_1 + j_2), h_1) + (v_0, \ell) \\ + (L(p)(j_2 - k), h_2 - \ell) + (G^{-1}(p)(j_1 + j_2), h_2) + (v_0, -\ell). \end{aligned}$$

The terms involving v_0 go out and we can collect to

$$\begin{aligned} (X(p)(j_1 + k), h_1 + \ell) + (G^{-1}(p)(j_1 + j_2), h_1 + h_2) \\ + (L(p)(j_2 - k), h_2 - \ell). \end{aligned} \quad (16)$$

This is the desired scalar product.

4.58 Let us now consider the $(n + m)$ -tuples

$$[a_1, a_2, \dots, a_n, b_1, \dots, b_m] = j_1 \oplus k \oplus j_2 \quad (17)$$

obtained from (2) by deleting the b_{m+1}, \dots, b_n . We still interpret these as currents into the relevant terminals of \mathbf{M}_{AD} . We also observe that when the current (17) is given, (2) can be determined, because by (10)

$$a_{m+s} + b_{m+s} = 0, \quad s = 1, 2, \dots, n - m.$$

Given (17), and therefore (2) or (10), we can determine the voltages (14) and (15), where v_0 is an arbitrary element of \mathbf{V}_1 . Let us agree now always so to choose v_0 that the component of (15) in the subspace \mathbf{V}_1 vanishes. This means that, in (17), we have specified arbitrarily the currents into the left-hand terminals of \mathbf{M}_{AD} (on A-A) and into the upper m of the right-hand terminals. We have also agreed that the voltages across the lower $n - m$ terminals on D-D shall be zero, so that (15) is an

n -tuple of the form

$$u \oplus 0 \quad (18)$$

where $u \in \mathbf{U}$. Regarding (15), with this determination of v_0 , as simply an m -tuple u (ignoring its last $n - m$ zero components), we see that (17) and the ordered pair (14), (15) are now currents and voltages in a $2(n + m)$ -pole \mathbf{M}_{AD}^* obtained from \mathbf{M}_{AD} by shorting and thereafter ignoring the lower $n - m$ terminals on D-D.

4.59 Now (17) is unrestricted. Given it, the corresponding voltages can be computed from (14) and (15) by determining v_0 so that (15) lies in \mathbf{U} . Hence \mathbf{M}_{AD}^* has an impedance matrix, since any single valued linear mapping from (17) to voltages can be described by a matrix. Our job is now to show that this matrix comes under 3.1. Before doing this, however, we shall point out that a realization of \mathbf{M}_{AD}^* provides one for \mathbf{M}_{AD} .

Fig. 7 shows how a $2(2n)$ -pole equivalent to \mathbf{M}_{AD} would be constructed from \mathbf{M}_{AD}^* . The equivalence is evident almost at once: The pairs of \mathbf{M}_{AD}^* are the currents (17) and the voltages (14) and (15) with a special determination of v_0 , where (15) is regarded as an m -tuple. The current (10) is clearly that which flows in the $2(2n)$ -pole of Fig. 7 when (17) flows in \mathbf{M}_{AD}^* . Furthermore, regarding (15) as an n -tuple of the form (18), we see that the voltages in Fig. 7 can be obtained from (14), (15) by adding an arbitrary voltage of the form

$$(0 \oplus v) \oplus (0 \oplus v),$$

where $v \in \mathbf{V}_1$ of course. This arbitrary added voltage eliminates the special role played by v_0 in (14) and (15). Hence therein v_0 itself may be considered to be an arbitrary element of \mathbf{V}_1 , and (14), (15) represent the voltages in Fig. 7. The pairs admitted by the $2(2n)$ -pole of Fig. 7 are therefore exactly those admitted by \mathbf{M}_{AD} , Q.E.D.

4.60 We have now established that \mathbf{M}_{AD}^* has an impedance matrix, say $M(p)$. $M(p)$ operates from an $(n + m)$ space of currents (17) of 4.58 to an $(n + m)$ space of voltages (14), (15) of 4.56, where in (15) we properly choose v_0 so that the last $(n - m)$ components are zero and can be ignored.

Now any impedance matrix $\hat{Z}(p)$ is completely determined when we know for each two currents m_1 and m_2 the scalar product

$$(\hat{Z}(p)m_1, m_2) \quad (1)$$

(Cf. Halmos⁹, par. 53). We shall make this computation for $M(p)$. The

currents (17) of 4.58 may be regarded as elements of the subspace (10) of 4.55. We have called this subspace \mathbf{K}_M^2 . The voltages (14), (15), with v_0 chosen to make (15) an n -tuple of the form (18) (4.58), are elements of a subspace \mathbf{V}_M^2 of \mathbf{V}^2 .

It is evident at once that the scalar product between a current $(n + m)$ -tuple (17) and the $(n + m)$ -tuple (14), (15) (v_0 properly chosen!) is exactly the same as the scalar product between the current $(2n)$ -tuple (10) and the $(2n)$ -tuple formed from the $(n + m)$ -tuple (14), (15) by adjoining $(n - m)$ zeros to expand (15) to an n -tuple of the form (18).

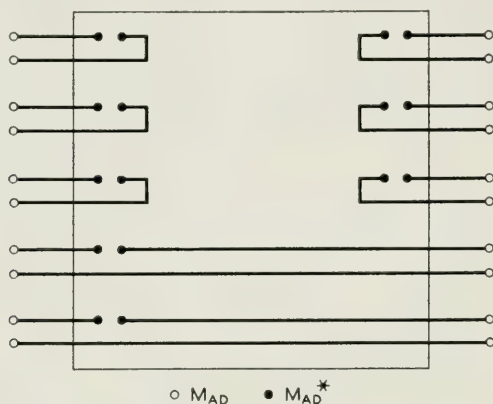


Fig. 7 = Construction of M_{AD} from M_{AD}^* . The solid terminals are those of M_{AD}^* , the open circles those of M_{AD} .

Now we know that we may regard (15) as an n -tuple of the form (18) by a suitable choice of v_0 . But we calculated in 4.57 the scalar product between an arbitrary $(2n)$ -tuple and (14), (15) with an arbitrary v_0 . The answer was (16) of 4.57. By proper choice of v_0 , then, (16) represents the bilinear form (1) above for $M(p)$. Since (16) is independent* of v_0 , it contains in itself the whole of the properties of $M(p)$.

4.61 To show that $M(p)$ is PR, we need show only that $M(p)$ is symmetric and that is *quadratic* form ($j_i = h_i$ and $k = \ell$ in (16)) is a PR function of p (2.09).

By their definitions, $X(p)$, $L(p)$, and $G^{-1}(p)$ are all symmetric. Hence if all currents are real, the value of (16) is unchanged by interchanging j_i with h_i , $i = 1, 2$, and k with ℓ . Therefore $M(p)$ is symmetric.

4.62 Henceforth we consider the quadratic from

* This is the gist of P3 of (I, 7.4). Use of the results of I here would have given a more direct but much less constructive representation of M_{AD} .

$$(X(p)(j_1 + k), j_1 + k) + (G^{-1}(p)(j_1 + j_2), j_1 + j_2) \\ + (L(p)(j_2 - k), j_2 - k) \quad (2)$$

obtained from (16). By the definitions of $X(p)$, $L(p)$, and $G^{-1}(p)$ this is a rational function taking real values for real p . Hence we need only show of (2) that its real part is non-negative when $\operatorname{Re}(p) > 0$ to show that it and $M(p)$ are PR.

Referring to (6) and (11) of paragraph 4.4 and (11) of 4.56 for the definitions, we see that (2) can be written

$$\frac{1}{p} \left[- (A(j_1 + k), j_1 + k) + \frac{\omega_0^2}{2} (G^{-1}(j_1 + j_2), j_1 + j_2) \right. \\ \left. + (H(j_2 - k), j_2 - k) \right] \\ + p \left[- (B(j_1 + k), j_1 + k) + \frac{1}{2} (G^{-1}(j_1 + j_2), j_1 + j_2) \right. \\ \left. + (F(j_2 - k), j_2 - k) \right]. \quad (3)$$

That is, the quadratic form in question has poles, simple ones, only at 0 and ∞ , and has no constant term. If we can show that the residues at these poles are non-negative, then it will follow not only that $M(p)$ is PR but that $M(p)$ is of the form

$$\frac{1}{p} M_0 + p M_\infty$$

where each of these summands is realizable by 3.1.

Unfortunately, there still remains some computation to verify that the residues of (3) are non-negative.

4.62 We first recapitulate some relations obtained earlier:

$$Z^{(2)}(p) = Z(p) + \frac{1}{p} A + pB; \quad (4)$$

this is (5) of 4.42.

$$Y^{(2)}(p) = \frac{2p}{p^2 + \omega_0^2} G + Y^{(3)}(p); \quad (5)$$

this is (7) and (8) of 4.45.

$$Z^{(3)}(p) = \frac{1}{p} H + pF + Z^{(4)}(p); \quad (6)$$

this is (11) and (12) of 4.491.

By their definitions,

$$Z^{(i)}(p) = [Y^{(i)}(p)]^{-1}$$

for $i = 2, 3$. By hypothesis, $Z(p)$ and

$$Z(p)^{-1} = Y(p)$$

are both finite everywhere on $p = i\omega$. By its construction, $Z^{(4)}(p)$ is finite at $p = 0$ and $p = \infty$.

4.63 We claim now that each $Y^{(i)}(p)$ is finite at $p = 0$ and ∞ , $i = 2, 3$.

Proof: We need consider only $Y^{(2)}(p)$ since $Y^{(3)}(p)$ differs from it by something which vanishes at $p = 0$ and $p = \infty$ ((5) above). Let

$$Y^{(2)}(p) = \tilde{Y}(p) + \frac{1}{p}E + pQ$$

where $\tilde{Y}(p)$ is finite at $p = 0$ and $p = \infty$. Since $Y^{(2)}(p)$ is PR (4.43), E and Q are real and symmetric.

Using the form (4) above for $Z^{(2)}(p)$,

$$\begin{aligned} 1 &= Z^{(2)}(p)Y^{(2)}(p) = Z(p)\tilde{Y}(p) + BE + AQ \\ &\quad + p(Z(p)Q + B\tilde{Y}(p)) + p^2BQ \\ &\quad + \frac{1}{p}(Z(p)E + A\tilde{Y}(p)) + \frac{1}{p^2}AE. \end{aligned} \quad (7)$$

Multiplying through by p^2 , p , $\frac{1}{p^2}$, $\frac{1}{p}$ and taking limits as $p \rightarrow 0, 0, \infty, \infty$, respectively, we obtain

$$\begin{aligned} AE &= 0 \\ Z(0)E + A\tilde{Y}(0) &= 0, \\ BQ &= 0, \\ Z(\infty)Q + B\tilde{Y}(\infty) &= 0. \end{aligned} \quad (8)$$

We can also write a formula like (7) with the factors in reverse order, and obtain the analogous forms to (8) in which the factors are commuted. Let us call these commuted relations (8'). Multiply the second relation (8) on the left by E and use the first relation of (8'). We obtain

$$EZ(0)E = 0. \quad (9)$$

Working similarly with the last two relations in (8) and (8'), we get

$$QZ(\infty)Q = 0. \quad (10)$$

Now let v be an arbitrary voltage in \mathbf{V} and let

$$w = Z(0)Ev.$$

Then $w \in \mathbf{V}$, and by (9) the current

$$Ew = 0$$

for any v . Hence

$$\begin{aligned} 0 &= (v, Ew) = \overline{(w, E^*v)} = (\bar{w}, \bar{E}^*\bar{v}) \\ &= (\bar{Z}(0)\bar{E}\bar{v}, E'\bar{v}) \end{aligned} \quad (11)$$

by (I, 7.2, 14.0). Now E is real and symmetric, as noted above. Hence $E = \bar{E} = E'$. Furthermore, $Z(0)$ is real, so (11) becomes

$$(Z(0)Eu, Eu) = 0 \quad (12)$$

where $u = \bar{v}$ is any element of \mathbf{V} . Now $Z(p)$ is non-singular on $p = i\omega$, and its real part is semidefinite there. At $p = 0$, $Z(0)$ is its own real part, hence semidefinite and non-singular, hence definite. Then (12) implies that $Eu = 0$. This being true for all $u \in \mathbf{V}$, $E = 0$.

The proof that $Q = 0$ follows similarly from (10).

4.64 With $Y^{(2)}(p)$ and $Y^{(3)}(p)$ simplified at $p = 0$ and ∞ , we can go back and compute

$$\begin{aligned} 1 &= Z^{(2)}(p)Y^{(2)}(p) \\ &= \left(Z(p) + \frac{1}{p}A + pB \right) \left(\frac{2p}{p^2 + \omega_0^2}G + Y^{(3)}(p) \right). \end{aligned} \quad (13)$$

Of the six terms obtained on expanding this exactly one, namely

$$\frac{1}{p}AY^{(3)}(p)$$

is not obviously finite at $p = 0$, and another,

$$pBY^{(3)}(p)$$

is not *a priori* finite at $p = \infty$. We conclude by multiplying through by p and letting $p \rightarrow 0$, and dually at $p = \infty$, that

$$\begin{aligned} AY^{(3)}(0) &= 0 = Y^{(3)}(0)A \\ BY^{(3)}(\infty) &= 0 = Y^{(3)}(\infty)B, \end{aligned} \quad (14)$$

where the commuted form can be established by a new calculation from $1 = Y^2(p)Z^2(p)$, or by taking transposes.

In a similar way, we compute from

$$1 = Z^{(3)}(p)Y^{(3)}(p) = Z^{(4)}(p)Y^{(3)}(p) + \frac{1}{p}HY^{(3)}(p) + pFY^{(3)}(p) \quad (15)$$

that

$$\begin{aligned} HY^{(3)}(0) &= 0 = Y^{(3)}(0)H, \\ FY^{(3)}(\infty) &= 0 = Y^{(3)}(\infty)F. \end{aligned} \quad (16)$$

Now $Y^{(3)}(p)$ is finite at 0 and ∞ , so we may expand it in a power series about either point. Let these be

$$\begin{aligned} Y^{(3)}(p) &= Y^{(3)}(0) + pY_1^{(3)}(0) + O(p^2), \\ Y^{(3)}(p) &= Y^{(3)}(\infty) + \frac{1}{p}Y_1^{(3)}(\infty) + O\left(\frac{1}{p^2}\right). \end{aligned} \quad (17)$$

Putting the appropriate one of these into (13) and taking a limit at 0 or ∞ we get, by using (14), that

$$\begin{aligned} 1 &= \frac{1}{\omega_0^2}AG + AY_1^{(3)}(0) + Z(0)Y^{(3)}(0), \\ 1 &= 2BG + BY_1^{(3)}(\infty) + Z(\infty)Y^{(3)}(\infty). \end{aligned} \quad (18)$$

A relation (18') with factors commuted is also true.

We may also put (17) into (15) and get

$$\begin{aligned} 1 &= Z^{(4)}(0)Y^{(3)}(0) + HY_1^{(3)}(0), \\ 1 &= Z^{(4)}(\infty)Y^{(3)}(\infty) + FY_1^{(3)}(\infty), \end{aligned} \quad (19)$$

and also a commuted form (19').

Right multiply the first line of (19) by A and the second by B , and use (14). This gives

$$\begin{aligned} A &= HY_1^{(3)}(0)A, \\ B &= FY_1^{(3)}(\infty)B. \end{aligned} \quad (20)$$

Left multiply the first line of (18') by H and the second by F . This gives, by (16),

$$\begin{aligned} H &= \frac{2}{\omega_0^2}HGA + HY_1^{(3)}(0)A, \\ F &= 2FGB + FY_1^{(3)}(\infty)B. \end{aligned} \quad (21)$$

Using (20) in (21), we have the relations

$$\frac{2}{\omega_0} HGA = H - A, \quad (22)$$

$$2FGB = F - B.$$

These are fundamental to the evaluation of the residues of (3). Before calculating these residues, we draw a further important conclusion from the formulas just developed.

Relation (20) exhibits A as a product of H and a possibly singular matrix (viz., $Y_1^{(3)}(0)A$). Hence

$$\text{rank}(A) \leq \text{rank}(H).$$

But relation (21) shows H as a product of A by

$$\frac{2}{\omega_0} HG + HY_1^{(3)}(0).$$

Hence

$$\text{rank}(H) \leq \text{rank}(A).$$

That is,

$$\begin{aligned} \text{rank}(A) &= \text{rank}(H), \\ \text{rank}(B) &= \text{rank}(F), \end{aligned} \quad (23)$$

the latter being established in the same way.

4.65 The formulas developed in 4.64 are all quite symmetric as between relations obtained at $p = \infty$ and those at $p = 0$. We shall now continue to the evaluation of the residue of (3) at $p = \infty$. The evaluation at $p = 0$ proceeds in an exactly similar manner.

The residue in question is, from (3),

$$\begin{aligned} -(B(j_1 + k), j_1 + k) + \frac{1}{2}(G^{-1}(j_1 + j_2), j_1 + j_2) \\ + (F(j_2 - k), j_2 - k). \end{aligned} \quad (24)$$

Here j_1 and j_2 are any elements of \mathbf{J} and k any element of \mathbf{K}_1 . The range of G is \mathbf{J} and the operator G^{-1} operates from \mathbf{J} to $\mathbf{U} = \mathbf{J}^*$, representing the inverse to the operation G from \mathbf{U} to \mathbf{J} .

Let us define h and eliminate j_2 by the relation

$$j_2 = 2h + 2GB(j_1 + k) - j_1. \quad (25)$$

Since the range of G is \mathbf{J} , $h \in \mathbf{J}$.

The definition analogous to (25) for the other pole of (3) is

$$j_2 = \frac{2}{\omega_0} h + \frac{2}{\omega_0} GA(j_1 + k) - j_1.$$

We shall now say no more about this pole.

Putting (25) into (24) we get at once the form

$$-(B(j_1 + k), j_1 + k) + (G^{-1}h + G^{-1}GB(j_1 + k), 2h + 2GB(j_1 + k)) \\ + (2Fh + 2FGB(j_1 + k) - Fj_1 - Fk, 2h + 2GB(j_1 + k) - j_1 - k).$$

Here we cannot at once put $G^{-1}G = 1$, because this is only true in \mathbf{U} . We expand in the following way: The first product is left intact, the second is expanded by distributivity into four terms, and in the third we use (22) and expand into five terms by distributivity. The ten resulting terms are:

$$-(B(j_1 + k), j_1 + k) + 2(G^{-1}h, h) \\ + 2(G^{-1}GB(j_1 + k), h) + 2(G^{-1}h, GB(j_1 + k)) \\ + 2(G^{-1}GB(j_1 + k), GB(j_1 + k)) \\ + 4(Fh, h) - 2(B(j_1 + k), h) \\ + 2(Fh, 2GB(j_1 + k) - j_1 - k) \\ - 2(B(j_1 + k), GB(j_1 + k)) + (B(j_1 + k), j_1 + k).$$

Enumerate these terms 1, 2, \dots , 10 in the order written. We shall show by combining that only 2 and 6 remain.

Clearly 1 and 10 cancel.

Consider the operator $G^{-1}G$ as we have defined it. If $v \in \mathbf{V}$, we can put

$$v = u + v_1$$

where $u \in \mathbf{U}$, $v_1 \in \mathbf{V}_1$. Then

$$Gv = Gu + Gv_1 = Gu,$$

because of the matrix form for G in the coordinate system chosen in 4.5. By definition of G^{-1} (in 4.56), since $u \in \mathbf{U}$,

$$G^{-1}Gu = u.$$

Hence, combining the last three relations,

$$G^{-1}Gv = v - v_1 \quad (26)$$

for any $v \in \mathbf{V}$, where v_1 is a suitable element of \mathbf{V}_1 (depending on v of course).

Using (26) in term 3, we get for this term

$$2(B(j_1 + k), h) - 2(v_1, h)$$

for some $v_1 \in \mathbf{V}_1$. But $h \in \mathbf{J} = (\mathbf{V}_1)^0$ ((1) of 4.5). Hence the second term here vanishes and term 3 cancels term 7. By an exactly similar argument, since $GB(j_1 + k) \in \mathbf{J}$, we find that term 5 cancels term 9.

Consider term 4, and write it in the form

$$\begin{aligned} 2(G^{-1}h, k_1) &= 2(\overline{(G^{-1})^*k_1}, h) \\ &= 2((\bar{G}^{-1})^*\bar{k}_1, \bar{h}) = 2(G^{-1}\bar{k}_1, \bar{h}). \end{aligned}$$

This follows by (I, 7.2, 14.0) and the fact that G^{-1} is symmetric. Putting in the definition of k_1 , and using the fact that G and B are real, we get

$$\begin{aligned} 2(G^{-1}\bar{k}_1, \bar{h}) &= 2(G^{-1}\bar{G}B(j_1 + \bar{k}), \bar{h}) \\ &= 2(G^{-1}GB(j_1 + \bar{k}), \bar{h}). \end{aligned}$$

Now \mathbf{J} is real (4.42) so $\bar{h} \in \mathbf{J}$. Therefore the reasoning used on term 3 yields finally

$$2(B(j_1 + \bar{k}), \bar{h})$$

as the value of term 4.

We now write term 8 as

$$2(Fh, k_2)$$

and transform it to

$$2(F\bar{k}_2, \bar{h}),$$

by the reasoning just used on 4. Putting in what k_2 is, this is

$$2(2F\bar{G}B(j_1 + \bar{k}) - Fj_1 - F\bar{k}, \bar{h}).$$

Using the reality of G and B , and (22), this is

$$-2(B(j_1 + \bar{k}), \bar{h}).$$

This cancels term 4 and all terms save 2 and 6 are accounted for. Finally, then, the residue of (3) at $p = \infty$ is

$$2(G^{-1}h, h) + 4(Fh, h). \quad (27)$$

Since G^{-1} is definite in \mathbf{J} and F is semidefinite, this residue is non-negative, and indeed not zero if $h \neq 0$ and $h \in \mathbf{J}$.

4.7 We have established the non-negativity of the residue of (3) at $p = \infty$. A similar argument (exactly parallel, in fact) will establish the same for the residue at $p = 0$. Each term in the representation

$$M(p) = \frac{1}{p} M_0 + p M_\infty$$

of 4.61 is then realizable by 3.1. Hence \mathbf{M}_{AD}^* is a realizable reactance $2(n + m)$ -pole, and so therefore is \mathbf{M}_{AD} , as we noted in discussing Figure 7 (4.59). Therefore, if $\bar{\mathbf{N}}^{(4)}$ of Figure 5 is physically realizable, so also is $\bar{\mathbf{N}}$ and therefore \mathbf{N} . We denote by \mathbf{N}_1 the $\bar{\mathbf{N}}^{(4)}$ obtained in this way from \mathbf{N} , and define IB as the operation which constructs \mathbf{N}_1 from \mathbf{N} .

We must still establish the claims made in 4.04 for IB. No properties of $\bar{\mathbf{N}}^{(4)} = \mathbf{N}_1$ have been proved beyond the existence of its impedance matrix, $Z^{(4)}(p)$, but this is all that is claimed in the third column of 4.04. The fifth column is also established. We must now however compare the degree of \mathbf{N}_1 , i.e., of $Z^{(4)}(p)$, with that of $Z(p)$.

By 2.13, 2.14 and 2.15 applied to (4), (5), and (6) of 4.62,

$$\begin{aligned}\delta(Z^{(2)}) &= \delta(Z) + \text{rank } (A) + \text{rank } (B), \\ \delta(Z^{(2)}) &= \delta(Y^{(2)}) = \delta(Y^{(3)}) + 2 \text{rank } (G), \\ \delta(Y^{(3)}) &= \delta(Z^{(3)}) = \delta(Z^{(4)}) + \text{rank } (H) + \text{rank } (F).\end{aligned}$$

We know $m = \text{rank } (G) \geq 1$. Let

$$r = \text{rank } (A) + \text{rank } (B).$$

Then from (23), and the relations above in order,

$$\begin{aligned}\delta(Z) &= \delta(Z^{(2)}) - r = (\delta(Z^{(3)}) + 2m) - r \\ &= (\delta(Z^{(4)}) + r) + 2m - r \\ &= \delta(Z^{(4)}) + 2m.\end{aligned}$$

Hence $\delta(Z) - \delta(Z^{(4)}) = \delta(\mathbf{N}) - \delta(\mathbf{N}_1) = 2m > 0$. The claims of 4.04 are then established.

4.71 We must yet verify 4.07 for IB. Let $\delta(M)$ be the degree of

$$M(p) = \frac{1}{p} M_0 + p M_\infty.$$

Then by 3.21, \mathbf{M}_{AD}^* , whose matrix is $M(p)$, can be realized with $\delta(M)$ reactive elements. By Figure 7, then \mathbf{M}_{AD} can be so realized, and it follows that exactly $\delta(M)$ reactive elements are comprised between \mathbf{N} and \mathbf{N}_1 under IB.

Now by 2.14 and 2.15,

$$\delta(M) = \text{rank } (M_0) + \text{rank } (M_\infty).$$

We shall compute the second term. The first is obtained by an exactly parallel calculation.

Using the fact that $M(p)$ is determined by its quadratic form, we see that M_∞ is the matrix whose form is the residue of that of $M(p)$ at $p = \infty$. This residue was computed in (27) of 4.65 to be

$$2(G^{-1}h, h) + 4(Fh, h) \quad (1)$$

when the current vector, (17) of 4.58, is

$$j_1 \oplus k \oplus j_2, \quad (2)$$

and, (25) of 4.65,

$$2h = j_1 + j_2 - 2GB(j_1 + k). \quad (3)$$

Here $j_1, j_2 \in \mathbf{J}$ and $k \in \mathbf{K}_1$.

Now M_∞ is an $(n + m) \times (n + m)$ matrix by construction. Then

$$\nu = n + m - \text{rank}(M_\infty) \quad (4)$$

is its nullity, the dimension of its null space. This is proved in Halmos⁹, par. 37, for dimensionless operators, and a similar proof applies to impedance operators.

Now for any symmetric and semidefinite impedance operator \hat{Z} , the null space of \hat{Z} is exactly the aggregate of currents k such that the quadratic form

$$(\hat{Z}k, k) = 0.$$

This may be seen at once by choosing a coordinate frame in which the matrix of \hat{Z} is diagonal. Since we know from 4.65 that M_∞ is symmetric and semidefinite, we can compute ν as the dimensionality of the space of vectors (2) above for which (1) vanishes.

As noted in 4.65, $h \in \mathbf{J}$, and (1) vanishes if and only if $h = 0$, because G^{-1} , as an operator from \mathbf{J} to \mathbf{U} , is definite (semidefinite and non-singular). Hence ν is the maximum number of linearly independent vectors (2) for which, from (3),

$$(1 - 2GB)j_1 + j_2 - 2GBk = 0. \quad (5)$$

The left member of (5) is a vector in \mathbf{J} depending linearly and homogeneously on the vector (2). Hence, regarding \mathbf{J} as a subspace of the space $\mathbf{J} \oplus \mathbf{K}_1 \oplus \mathbf{J}$ in which (2) lies, the left member of (5) is the value in $\mathbf{J} \oplus \mathbf{K}_1 \oplus \mathbf{J}$ of a certain linear operation applied to the vector (2). Let us call this operator P . The number ν , by definition the number of linearly independent vectors (2) for which (5) holds, is the nullity of P . The dimension of P is $n + m$, and its rank is clearly m because the left member of (5)—a typical element in the range of P —lies in \mathbf{J} and by

suitable choice of j_2 can be made to be any element of \mathbf{J} . Hence the nullity of P is $(n + m) - m = n$ (Halmos⁹, par. 37). That is

$$\nu = n,$$

and, by (4)

$$\text{rank } (M_\infty) = m.$$

A parallel argument will establish the same result for M_0 . Hence

$$\delta(M) = 2m = \delta(\mathbf{N}) - \delta(\mathbf{N}_1)$$

by a result of 4.7. Therefore \mathbf{M}_{AD}^* and \mathbf{M}_{AD} can be realized with

$$\delta(\mathbf{N}) - \delta(\mathbf{N}_1)$$

reactive elements and 4.07 holds for IB.

V. THE DEGREE OF A RATIONAL MATRIX

5.0 In this section we consider arbitrary $n \times n$ matrices $Z(p)$ whose elements are rational functions of the complex variable p . They are treated, generally, as arrays of functions with certain rules for addition, multiplication, and reciprocation, without geometric interpretation. A geometric development is possible, but would be cumbersome. Related ideas may be found, geometrically developed, in Appendix I of Halmos⁹.

This section deals wholly with concepts well known in the algebraic theory of matrices over an arbitrary field—in this case the field of rational functions. I have not found, however, any place where the particular developments which seem to be needed here are made sufficiently explicitly for reference. Accordingly, the presentation here is somewhat detailed. The particular path of argument followed is only one of many possible; it was chosen to lead easily to results needed in Section 6, and to parallel generally the rest of the paper.

This section could be abbreviated somewhat if one restricted himself to PR matrices $Z(p)$. We prefer not to limit the applicability of these results, however, since they may well be useful in non-passive realizability theory.

5.01 *Definition:* If $R(p)$ is a rational function of the form

$$R(p) = (p - p_0)^m R_1(p),$$

where $R_1(p)$ is finite and not zero at p_0 , and m may be of any sign, we call m the exponent of $(p - p_0)$ in $R(p)$. The number

$$r = \sup (-m, 0)$$

is called the order of the pole of $R(p)$ at p_0 , even if $r = 0$.

5.1 Let $Z(p)$ be an $n \times n$ matrix whose elements $Z_{rs}(p)$ are rational functions of the complex variable p . We write

$$Z_{rs}(p) = \frac{N_{rs}(p)}{D_{rs}(p)},$$

where N_{rs} and D_{rs} are relatively prime polynomials. Let $\Psi_Z(p)$ be the least common multiple of all $D_{rs}(p)$, ($1 \leq r, s \leq n$), so normalized that the coefficient of the highest power of p appearing in $\Psi_Z(p)$, (the *leading coefficient*) is unity. Then $\Psi_Z(p)$ is uniquely determined by $Z(p)$.

The matrix $\Psi_Z(p)Z(p)$ has polynomial elements. Its *Smith normal form* is a diagonal matrix $E(p)$,

$$E(p) = \begin{bmatrix} E_1(p) & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & E_2(p) & & & & \\ \cdot & & \cdot & & & \\ \cdot & & & \cdot & & \\ \cdot & & & & E_R(p) & \\ 0 & & & & 0 & \cdot \\ & & & & & \cdot & 0 \end{bmatrix} = A(p)\Psi_Z(p)Z(p)B(p), \quad (1)$$

with the following properties:

- (a) R is the rank of $\Psi_Z(p)Z(p)$.
- (b) Each $E_i(p)$, $1 \leq i \leq R$, is a polynomial with unit leading coefficient.
- (c) Each $E_i(p)$ is a factor of $E_{i+1}(p)$, $1 \leq i \leq R - 1$.
- (d) $A(p)$ and $B(p)$ are polynomial matrices, each with a constant non-vanishing determinant.
- (e) $E_1(p)E_2(p) \cdots E_k(p)$ is the normalized (and therefore unique) highest common factor of all k -rowed minor determinants of $\Psi_Z(p)Z(p)$.

These properties of $E(p)$ are developed for example, in Bocher¹⁵, Theorems 2 and 3 of paragraph 91 and Theorem 1 of paragraph 94. A simple variation of this last cited theorem will also prove the following uniqueness lemma.

5.11 *Lemma:* If some $E^0(p)$ satisfies (1) and (a), (b), (c) and (d) above, all written with superscripts on each E , and on A and B , then $E^0(p) = E(p)$.

Proof: $E^0(p)$ is equivalent to $E(p)$ in the sense of paragraph 94 of

Bocher¹⁵, for

$$E^0(p) = A^0(p)A^{-1}(p)E(p)B^{-1}(p)B^0(p).$$

Therefore it is also equivalent in the sense of par. 91 of Bocher¹⁵, (for this is Theorem 1 of paragraph 94). Hence the normalized greatest common factor of all k -rowed minors of $E^0(p)$ is the same as that of $E(p)$, that is, $E_1(p) \cdots E_k(p)$. But the greatest common factor of all k rowed minors of $E^0(p)$ is $E_1^0(p) \cdots E_k^0(p)$, because of property (c). In particular then $E_1(p) = E_1^0(p)$, and consequently $E_k(p) = E_k^0(p)$ by induction for $1 \leq k \leq R$. Q.E.D.

5.12 *Definition*: The normal form $W(p)$ of $Z(p)$ is the matrix $\Psi_z^{-1}(p)E(p)$. We write the elements of $W(p)$ in their lowest terms,

$$W(p) = A(p)Z(p)B(p) = \begin{bmatrix} \frac{e_1(p)}{\Psi_1(p)} & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & \frac{e_2(p)}{\Psi_2(p)} & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \frac{e_R(p)}{\Psi_R(p)} & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 0 & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & 0 \end{bmatrix} \quad (2)$$

with the polynomials $e_k(p)$, $\Psi_k(p)$ each having unit leading coefficients.

5.13 *Theorem*: The normal form $W(p)$ of $Z(p)$, as given by (2), has the properties (a'), (b'), (c'), (d'), and (e') listed below. Furthermore, any $W^0(p)$, given by (2) with superscripts on W , A , B , e_k , and Ψ_k ($1 \leq k \leq R$), which satisfies (a'), (b'), (c'), and (d') with corresponding superscripts, is in fact $W(p)$.

(a') R is the rank of $Z(p)$

(b') For each k , $1 \leq k \leq R$, $e_k(p)$ and $\Psi_k(p)$ are relatively prime polynomials with unit leading coefficients.

(c') Each $e_k(p)$ is a factor of $e_{k+1}(p)$, $1 \leq k \leq R - 1$, and each $\Psi_j(p)$ is a factor of $\Psi_{j-1}(p)$, $2 \leq j \leq R$.

(d') $A(p)$ and $B(p)$ are polynomial matrices each with a constant non-vanishing determinant

(e') $\Psi_1(p) = \Psi_z(p)$.

Proof: (a') and (d') follow immediately from (a) and (d) of 5.1. (b') is a matter of definition. (c') follows from (c) of 5.1 and the definition, 5.12, since the effect of cancelling common factors in each fraction of the sequence

$$\frac{E_1(p)}{\Psi_z(p)}, \frac{E_2(p)}{\Psi_z(p)}, \cdots, \frac{E_R(p)}{\Psi_z(p)}$$

cannot remove from any $E_k(p)$ a factor which was present in earlier $E_j(p)$ ($j < k$) but was not cancelled therefrom (treat each linear factor of Ψ_z and of E_1 as distinct, and each linear factor of $\frac{E_{k+1}(p)}{E_k(p)}$ as distinct to see this easily).

Property (e') is best proved by a reductio ad absurdum. We recall that $E_1(p)$ is the highest common factor of all elements of $\Psi_z(p)Z(p)$. Suppose now that $E_1(p)$ contained a factor φ in common with $\Psi_z(p)$. Then every non-zero element of $\Psi_z(p)Z(p)$ contains the factor φ . Hence no denominator in $Z(p)$ cancels φ from $\Psi_z(p)$. Hence no denominator contains φ as a factor, but this denies its presence in their least common multiple, $\Psi_z(p)$.

The uniqueness of $W(p)$ follows at once from the uniqueness lemma, 5.11. Multiply (2) by $\Psi_z(p)$. Then

$$\Psi_z(p)W^0(p) = A^0(p)\Psi_z(p)Z(p)B^0(p) \quad (3)$$

has diagonal elements of the form

$$\frac{\Psi_z(p)e_k(p)}{\Psi_k(p)}, \quad 1 \leq k \leq R. \quad (4)$$

But by (3) and (d'), these are the result of polynomial operations on the polynomial matrix $\Psi_z(p)Z(p)$. Hence the elements (4) are polynomials, and each has unit leading coefficient. $\Psi_z(p)W^0(p)$ then clearly satisfies (a), (b), (c), and (d) of 5.1. Therefore by 5.11, $\Psi_z(p)W^0(p) = E(p) = \Psi_z(p)W(p)$. Therefore $W^0(p) = W(p)$. Q.E.D.

5.14 *Corollary:* $W(p)$ is its own normal form.

5.15 *Corollary:* Let $\varphi(p)$ be a rational function and

$$Z_1(p) = \varphi(p)Z(p).$$

Let $W(p)$ be the normal form of $Z(p)$ and $W_1(p)$ the normal form of $Z_1(p)$. Then, when written in normalized lowest terms,

$$W_1(p) = \varphi(p)W(p).$$

Proof: Supposing that (2) above holds for W and Z , we have

$$\varphi(p)W(p) = A(p)Z_1(p)B(p).$$

Call the left side of this equation $W_1^0(p)$. We must identify this with $W_1(p)$. We have just showed that it satisfies (d') of 5.13. It clearly satisfies (a'), (b') and (c'), with Z_1 written for Z . Hence 5.13 implies the desired equality.

5.16 *Corollary:* If $C(p)$ and $D(p)$ are polynomial matrices with constant non-vanishing determinants, then the normal forms of $Z(p)$ and $C(p)Z(p)D(p)$ are the same.

Proof:

$$AZB = (AC^{-1})CZD(D^{-1}B)$$

and the bracketed factors are again polynomial matrices with constant non-vanishing determinants.

5.2 *Definition:* The point p_0 is a pole of $Z(p)$ if some element of $Z(p)$ has a pole at $p = p_0$. If p_0 is not a pole of $Z(p)$, we say that $Z(p_0)$ is finite, or that $Z(p)$ is finite at p_0 .

5.21 If p_0 is a pole of $Z(p)$, we may expand each element of Z in partial fractions and collect those terms having poles at p_0 , obtaining, when $p_0 \neq \infty$,

$$Z(p) = (p - p_0)^{-r}Z_r + (p - p_0)^{-r+1}Z_{r-1} + \cdots + (p - p_0)^{-1}Z_1 + Z_0(p), \quad (1)$$

where $Z_0(p_0)$ is finite, $Z_r \neq 0$, and the Z_k , $1 \leq k \leq r$, are matrices of constants. If $p_0 = \infty$, we read p^ℓ for $(p - p_0)^{-\ell}$ in (1), $1 \leq \ell \leq r$. All of $Z_0(p)$, Z_1 , \cdots , Z_r are uniquely defined by their construction from $Z(p)$.

5.22 *Definition:* If $Z(p)$ is given by (1) above, then r is the order of the pole of $Z(p)$ at p_0 .

5.23 Clearly, if $Z(p)$ has the form (1) at $p_0 \neq \infty$, some non-vanishing element of $Z(p)$ has a denominator containing the factor $(p - p_0)^r$, and no element has a pole of order higher than r at p_0 . Hence $(p - p_0)^r$ divides $\Psi_Z(p)$, but no higher power of $(p - p_0)$ does. Therefore, by (e') of 5.13, the normal form $W(p)$ of $Z(p)$ has a first element with an r^{th} order pole at p_0 . In particular, then, $p_0 \neq \infty$ is a pole of order r of $Z(p)$ if and only if it is a pole of order r of $W(p)$.

5.24 *Definition:* Consider a pole of order r of $Z(p)$, say p_0 , with $p_0 \neq \infty$. In the normal form $W(p)$ of $Z(p)$, (2) of 5.12, let γ_k be the order of the pole of the k^{th} diagonal element

$$\frac{e_k(p)}{\Psi_k(p)}$$

at the point $p = p_0$. Then $\gamma_k \geq \gamma_{k+1}$, and $\gamma_1 = r$. We write the γ_k in an ordered array

$$S(Z, p_0) = [\gamma_1, \gamma_2, \cdots, \gamma_n].$$

5.25 *Definition:* Consider two matrices $Z(p)$ and $Z_1(p)$, with

$$\begin{aligned} S(Z, p_0) &= [\gamma_1, \gamma_2, \dots, \gamma_n], \\ S(Z_1, p_0) &= [\gamma'_1, \gamma'_2, \dots, \gamma'_n]. \end{aligned}$$

We say

$$S(Z, p_0) \geq S(Z_1, p_0) \quad (2)$$

if and only if

$$\gamma_1 + \gamma_2 + \dots + \gamma_k \geq \gamma'_1 + \gamma'_2 + \dots + \gamma'_k$$

for every $k = 1, 2, \dots, n$. We say

$$S(Z, p_0) = S(Z_1, p_0) \quad (3)$$

if

$$\gamma_k = \gamma'_k$$

for $k = 1, 2, \dots, n$. It is easy to see that (3) is equivalent to the simultaneous validity of (2) and the reverse inequality.

5.26 *Theorem:* Let $p_0 \neq \infty$ be a pole of $Z(p)$. Let $F(p)$ be a rational $n \times n$ matrix which is finite at p_0 . Then

$$S(Z, p_0) \geq S(FZ, p_0).$$

In particular, if $F(p)$ is also non-singular at p_0 , then

$$S(Z, p_0) = S(FZ, p_0).$$

Proof: Let $\psi_F(p)$ and $\psi_Z(p)$ be the least common denominators of the elements of $F(p)$ and $Z(p)$, respectively. Then the exponent of $(p - p_0)$ in $\psi_Z(p)$ is r , while in $\psi_F(p)$ it is zero by 5.23.

Let $-\varepsilon_k$ be the exponent of $(p - p_0)$ in the k^{th} diagonal element of the normal form of Z , and $-\varepsilon'_k$ the similar quantity for FZ . Then

$$\begin{aligned} \varepsilon_1 \geq \varepsilon_2 \geq \dots \geq \varepsilon_n, \\ \varepsilon'_1 \geq \varepsilon'_2 \geq \dots \geq \varepsilon'_n, \end{aligned} \quad (3)$$

by (c') of 5.13. Let

$$\begin{aligned} \gamma_k &= \sup (\varepsilon_k, 0), \\ \gamma'_k &= \sup (\varepsilon'_k, 0), \end{aligned}$$

Then $\gamma_k \geq \varepsilon_k$, $\gamma'_k \geq \varepsilon'_k$, and

$$\begin{aligned} S(Z, p_0) &= [\gamma_1, \gamma_2, \dots, \gamma_n], \\ S(FZ, p_0) &= [\gamma'_1, \gamma'_2, \dots, \gamma'_n]. \end{aligned}$$

By 5.15, the normal form of FZ is

$$(\psi_F \psi_Z)^{-1} \cdot (\text{normal form of } \psi_F \psi_Z FZ).$$

Hence the exponent of $(p - p_0)$ in the k^{th} diagonal element of the normal form of $\psi_F \psi_Z FZ$ is $r - \varepsilon'_k$. By a similar argument, the exponent of $(p - p_0)$ in the k^{th} diagonal element of the normal form of $\psi_Z Z$ is $r - \varepsilon_k$. Hence, by (e) of 5.1,

$$(r - \varepsilon'_1) + \cdots + (r - \varepsilon'_b)$$

is the exponent of $(p - p_0)$ in the highest common factor of all b -rowed minor determinants of $\psi_F \psi_Z FZ$. Similarly

$$(r - \varepsilon_1) + \cdots + (r - \varepsilon_b)$$

is the exponent of $(p - p_0)$ in the highest common factor of all b -rowed minor determinants of $\psi_Z Z$.

Now $\psi_F \psi_Z FZ$ is a polynomial matrix. A typical b -rowed minor determinant of this matrix is of the form

$$\psi_F^b \psi_Z^b \sum M_b N_b, \quad (4)$$

where the summation is over certain products $M_b N_b$ of b -rowed minors M_b of F and b -rowed minors N_b of Z . For a proof of this, see MacDuffee¹⁶, Theorem 99.1. The expression (4) is the same as

$$\sum (\psi_F^b M_b)(\psi_Z^b N_b) \quad (5)$$

where the factors $(\psi_Z^b N_b)$ are now b -rowed minors of $\psi_Z Z$. If φ is a factor common to all b -rowed minors of $\psi_Z Z$, it certainly is a factor common to all expressions (4) or (5). Hence the highest common factor of all b -rowed minor determinants of $\psi_F \psi_Z FZ$ —i.e., of all expressions (4) or (5),—has an exponent for $(p - p_0)$ no lower than that in the highest common factor of all b -rowed minor determinants of $\psi_Z Z$. Hence for any b ,

$$(r - \varepsilon'_1) + \cdots + (r - \varepsilon'_b) \geq (r - \varepsilon_1) + \cdots + (r - \varepsilon_b),$$

or

$$\varepsilon_1 + \cdots + \varepsilon_b \geq \varepsilon'_1 + \cdots + \varepsilon'_b.$$

It follows that

$$\gamma_1 + \cdots + \gamma_b \geq \varepsilon'_1 + \cdots + \varepsilon'_b.$$

This being true for every b , it is certainly true for every b such that all terms on the right are ≥ 0 (cf. (2)). This means that for $b = 1$, and for

every successive $b > 1$ such that $\varepsilon'_b \geq 0$,

$$\gamma_1 + \cdots + \gamma_b \geq \gamma'_1 + \cdots + \gamma'_b.$$

This inequality is now not altered if non-negative numbers are added to its left member and zeros to its right member. Hence it holds for all b , $1 \leq b \leq n$, and

$$S(Z, p_0) \geq S(FZ, p_0). \quad (6)$$

This is the first claim of the theorem.

Now if $F(p)$ is non-singular at p_0 , then $F^{-1}(p)$ is rational, and finite at p_0 . Hence by what is already proved,

$$S(FZ, p_0) \geq S(F^{-1}(FZ), p_0).$$

This last array is just $S(Z, p_0)$. Hence we have (6) and its reverse, and the theorem is proved.

5.27 *Theorem*: If $p_0 \neq \infty$ and

$$Z(p) = Z_1(p) + Z_2(p),$$

where $Z_2(p)$ is finite at p_0 , then

$$S(Z, p_0) = S(Z_1, p_0).$$

The proof of this depends upon the following lemma.

5.28 *Lemma*: Let $Z^*(p)$ be such that at $p = p_0 \neq \infty$ its only elements having poles are on the main diagonal. Let $-\varepsilon'_1, -\varepsilon'_2, \cdots$ be the exponents of $(p - p_0)$ in the diagonal elements of $Z^*(p)$, so enumerated that

$$+\varepsilon'_1 \geq +\varepsilon'_2 \geq \cdots \geq +\varepsilon'_n.$$

Let $-\varepsilon_1, -\varepsilon_2, \cdots, -\varepsilon_n$ be the exponents of $(p - p_0)$ in the successive diagonal elements of the normal form of $Z^*(p)$. Then if $\varepsilon'_b \geq 0$ we have

$$\varepsilon'_1 + \cdots + \varepsilon'_b \geq \varepsilon_1 + \cdots + \varepsilon_b.$$

Proof: There exist constant non-singular matrices F, G such that FZ^*G has the same rows and columns as Z^* so permuted that the diagonal elements of FZ^*G are arranged in the order of ascending powers of $(p - p_0)$, the highest order pole being in the first position. Since the normal forms of Z^* and FZ^*G are identical, it suffices to consider Z^* itself to be in this form.

Let $\psi = \psi_{Z^*}(p)$. Now ψZ^* has its diagonal elements in the order of increasing positive power of $(p - p_0)$. Furthermore, any off-diagonal element of ψZ^* has $(p - p_0)^r$ as a factor.

Let b be such that $\varepsilon'_b \geq 0$. Any b -rowed minor of ψZ^* is a sum of products of b elements of ψZ^* . That b -rowed minor which has in it a term with a lowest possible exponent of $(p - p_0)$ is the upper left b -rowed minor. Even this minor has a term with exponent

$$(r - \varepsilon'_1) + \cdots + (r - \varepsilon'_b) \quad (7)$$

for $(p - p_0)$, this term being the product of the main diagonal elements. Hence the highest common factor of all b -rowed minors of ψZ^* has an exponent for $(p - p_0)$ not less than (7). Hence

$$(r - \varepsilon_1) + \cdots + (r - \varepsilon_b) \quad (8)$$

is not less than (7), since this is the exponent of $(p - p_0)$ in the product of the first b diagonal elements of the normal form of ψZ^* . The inequality between (8) and (7) is just the conclusion claimed in the lemma.

5.281 *Proof of 5.27:* Let

$$W(p) = A(p)Z(p)B(p)$$

be the normal form of $Z(p)$. Then

$$W = AZ_1B + AZ_2B. \quad (9)$$

If we expand all three terms here in Laurent series about p_0 , the term AZ_2B contributes no negative powers. It follows then from the diagonal form of W that the matrix

$$Z^* = AZ_1B$$

satisfies the conditions of 5.28. The ε'_k of that lemma are, from (9), just the exponents of $(p - p_0)$ in the successive diagonal elements of W , the normal form of Z , and the ε_k of 5.28 are the similar quantities for the normal form of $Z^* = AZ_1B$. But the normal form of AZ_1B is the same as that of Z_1 (5.16). Therefore in the inequality of 5.28 we may interpret all of the ε 's as exponents in the respective normal forms of Z and Z_1 .

Now

$$Z_1(p) = Z(p) + (-Z_2(p))$$

and $-Z_2(p)$ is again finite at p_0 . Hence we may conclude by the argument just used that if $\varepsilon_b \geq 0$ also

$$\varepsilon_1 + \cdots + \varepsilon_b \geq \varepsilon'_1 + \cdots + \varepsilon'_b.$$

Hence if either of ε_b or ε'_b is non-negative

$$\varepsilon_1 + \cdots + \varepsilon_b \geq \varepsilon'_1 + \cdots + \varepsilon'_b.$$

By induction on b , then,

$$\varepsilon_k = \varepsilon'_k$$

for $k = 1, 2$, etc. until such k that both are negative. Therefore

$$\gamma_k = \sup (\varepsilon_k, 0) = \gamma'_k = \sup (\varepsilon'_k, 0)$$

for all $k = 1, 2, \dots, n$. That is,

$$S(Z, p_0) = S(Z_1, p_0),$$

Q.E.D.

5.29 *Theorem*: Let $Z(p)$ be such that at $p = p_0 \neq \infty$ its only elements having poles lie on the main diagonal. Let $\sigma_1, \sigma_2, \dots, \sigma_n$ be the orders of these poles, so enumerated that

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n.$$

Then

$$S(Z, p_0) = [\sigma_1, \sigma_2, \dots, \sigma_n].$$

Proof: We write

$$Z(p) = Z^*(p) + Z_2(p),$$

where $Z^*(p)$ is diagonal, having exactly the diagonal elements of $Z(p)$. By 5.27,

$$S(Z, p_0) = S(Z^*, p_0).$$

Now $Z^*(p)$ falls under 5.28, but is diagonal in addition. In the proof of 5.28, therefore, it is exactly the principal minors of ψZ^* which have the lowest exponents for $(p - p_0)$, since all non-principal minors vanish and have zeros of arbitrary order at $p = p_0$. Furthermore, (7) is exactly the least exponent of $(p - p_0)$ in any b -rowed minor of ψZ^* since the principal minors are simple products. Hence (7) and (8) are equal, for any $b = 1, 2, \dots, n$. Therefore the exponents in the normal form of Z^* are exactly those of Z^* and

$$S(Z, p_0) = S(Z^*, p_0) = [\sigma_1, \sigma_2, \dots, \sigma_n].$$

Q.E.D.

5.3. *Definition*: Let

$$p = T(q) = \frac{\alpha q + \beta}{\gamma q + \delta}$$

be a non-singular bi-rational transformation from the q -sphere to the

p -sphere. Denote its inverse by

$$q = T^{-1}(p).$$

Given a rational $Z(p)$ the matrix

$$Z_1(q) = Z(T(q))$$

is rational in q .

For any p_0 such that $T^{-1}(p_0) \neq \infty$, we define

$$S_T(Z, p_0) = S(Z_1, T^{-1}(p_0)).$$

5.31 *Theorem*: If p_0 and $T^{-1}(p_0)$ are both finite,

$$S_T(Z, p_0) = S(Z, p_0).$$

Proof: Let $W_1(q)$ be the normal form of

$$Z_1(q) = Z(T(q)).$$

We have

$$W_1(q) = A(q)Z_1(q)B(q).$$

Consider

$$W_2(p) = W_1(T^{-1}(p)) = A(T^{-1}(p))Z(p)B(T^{-1}(p)).$$

Here the pre- and post factors of $Z(p)$ are rational, finite, and non-singular at p_0 . Hence by 5.26

$$S(W_2, p_0) = S(Z, p_0). \quad (1)$$

Let $q_0 = T^{-1}(p_0)$. It is then easily computed that the inverse transformation $T^{-1}(p)$ takes the form

$$q - q_0 = \frac{a(p - p_0)}{b(p - p_0) + 1}, \quad a \neq 0.$$

Any given diagonal element of $W_1(q)$ is of the form

$$(q - q_0)^\epsilon R(q),$$

where ϵ may have any sign, and $R(q)$ is rational, finite, and not zero at q_0 . The corresponding diagonal element of $W_2(p)$ is then

$$(p - p_0)^\epsilon \left(\frac{a}{b(p - p_0) + 1} \right)^\epsilon R_1(p),$$

where $R_1(p) = R(T^{-1}(p))$, and the factor multiplying $(p - p_0)^\epsilon$ is again

finite and not zero at p_0 . The exponents of $(p - p_0)$ in the elements of $W_2(p)$ are therefore exactly the exponents of $q - q_0$ in the elements of $W_1(q)$. From 5.29, then

$$S(W_2, p_0) = S(W_1, q_0).$$

This with (1) and the definition 5.3 proves the theorem.

5.32 Definition: Given any p , let $p = T(q)$ be a non-singular bi-rational transformation such that $q_0 = T^{-1}(p_0) \neq \infty$. We define $S^*(Z, p_0)$ by

$$S^*(Z, p_0) = S_T(Z, p_0).$$

5.33 Lemma: $S^*(Z, p_0)$ is independent of the T chosen to define it.

Proof: Consider $q = T^{-1}(p)$ and $r = U^{-1}(p)$, each such that p_0 is mapped on a finite point. Then by definition

$$S_T(Z, p_0) = S(Z_1, q_0),$$

$$S_U(Z, p_0) = S(Z_2, r_0),$$

where

$$q_0 = T^{-1}(p_0), \quad r_0 = U^{-1}(p_0),$$

$$Z_1(q) = Z(T(q)),$$

$$Z_2(r) = Z(U(r)).$$

Now $r = U^{-1}(T(q)) = V(q)$, say, and r_0 and q_0 are finite. Hence by 5.31

$$S_V(Z_2, r_0) = S(Z_2, r_0) = S_U(Z, p_0). \quad (2)$$

But by definition

$$S_V(Z_2, r_0) = S(Z_3, V^{-1}(r_0)) = S(Z_3, q_0) \quad (3)$$

where

$$Z_3(q) = Z_2(V(q))$$

But

$$Z_2(V(q)) = Z(U(U^{-1}(T(q)))) = Z(T(q)) = Z_1(q).$$

Hence

$$S(Z_3, q_0) = S(Z_1, q_0) = S_T(Z, p_0).$$

This, with (2) and (3), proves the lemma.

5.34 Theorem: Theorems 5.26, 5.27, and 5.29 hold for S^* without the restriction that p_0 be finite.

Proof: Let $q_0 = T^{-1}(p_0) \neq \infty$. For 5.26 we have

$$S^*(Z, p_0) = S(Z_1, q_0) \geq S(F_1 Z_1, q_0) = S^*(FZ, p_0)$$

where the equalities are by definition and the inequality is 5.26 applied to matrices rational in q , since

$$F_1(q) = F(T(q))$$

is by hypothesis finite at q_0 . The remaining conclusion of 5.26 follows similarly. The proofs of 5.27 and 5.29 are equally simple.

5.35 *Theorem:* If we extend 5.3 to S^* by defining

$$S_T^*(Z, p_0) = S^*(Z_1, T^{-1}(p_0)),$$

then 5.31 holds for S^* with no restrictions on p_0 or $T^{-1}(p_0)$.

Proof: By their definitions,

$$S_T^*(Z, p_0) = S^*(Z_1, T^{-1}(p_0)) = S_U(Z_1, T^{-1}(p_0)), \quad (4)$$

where U is such that $U^{-1}(T^{-1}(p_0))$ is finite. But

$$S_U(Z_1, T^{-1}(p_0)) = S(Z_2, U^{-1}(T^{-1}(p_0))) \quad (5)$$

where

$$Z_2(r) = Z_1(U(r)) = Z(T(U(r))).$$

Let $V(r) = T(U(r))$. Then, by definitions,

$$S(Z_2, U^{-1}(T^{-1}(p_0))) = S_V(Z, p_0) = S^*(Z, p_0), \quad (6)$$

since $V^{-1}(p_0) = U^{-1}(T^{-1}(p_0))$ is finite. The theorem follows from (4), (5), and (6).

5.4 *Definition:* Let

$$S^*(Z, p_0) = [\gamma_1, \gamma_2, \dots, \gamma_n].$$

Define

$$\delta(Z, p_0) = \gamma_1 + \gamma_2 + \dots + \gamma_n,$$

$$\delta(Z) = \sum \delta(Z, p_0),$$

where the latter summation is over all poles p_0 of $Z(p)$, including $p_0 = \infty$. This $\delta(Z)$ is the degree of Z for which we must establish the properties claimed in 2.11 through 2.17. These properties will be demonstrated in 5.41 through 5.45, in numerical order, saving 2.13, which is deferred to 5.46.

5.41 Clearly $\delta(Z)$ is an integer and non-negative. If $\delta(Z) = 0$, then every γ at every p_0 is zero. Hence no p_0 , not even ∞ , is a pole of Z . Hence each element of $Z(p)$ is a constant. This establishes 2.11 and 2.12.

5.42 Suppose

$$Z(p) = Z_1(p) + Z_2(p)$$

where each $Z_i(p)$ is finite at every pole of the other. The poles of $Z(p)$ are then exactly the poles $p_0^{(1)}$ of Z_1 and those $p_0^{(2)}$ of Z_2 . At each pole, 5.27 applies in the enlarged sense of 5.34, so

$$\delta(Z, p_0^{(i)}) = \delta(Z_i, p_0^{(i)}).$$

Breaking the sum defining $\delta(Z)$ into sums over the $p_0^{(1)}$ and $p_0^{(2)}$ proves that

$$\delta(Z) = \delta(Z_1) + \delta(Z_2).$$

This is 2.14.

5.43 If

$$Z(p) = f(p)R,$$

where R is a constant matrix, then the normal form of $Z(p)$ is $f(p)$ times a diagonal matrix of the same rank as R (5.15). 2.15 then follows at once.

5.44 If

$$Z_1(p) = AZ(p)B,$$

where A and B are constant and non-singular, the poles of $Z_1(p)$ and $Z(p)$ are the same. At each, 5.26 applies in the enlarged sense of 5.34. Therefore $\delta(Z_1) = \delta(Z)$. This is 2.16.

5.45 If $Z_1(p)$ is $Z(p)$ bordered by zeros, they have the same poles. One verifies at once from 5.11 that the normal form of $Z_1(p)$ is that of $Z(p)$ bordered by zeros. Since also $Z_1(T(q))$ is $Z(T(q))$ bordered by zeros, it follows that

$$S^*(Z_1, p_0) = S^*(Z, p_0)$$

at every pole, whence $\delta(Z_1) = \delta(Z)$. This is 2.17.

5.46 We must prove that if $Z(p)$ is non-singular, then

$$\delta(Z) = \delta(Z^{-1})$$

Proof: Choose a bi-rational transformation $p = T(q)$ such that at

$p = T(\infty)$ both of $Z(p)$ and $Z^{-1}(p)$ are finite. Let

$$Z_1(q) = Z(T(q)).$$

Then

$$Z_1^{-1}(q) = Z^{-1}(T(q)).$$

Let $W_1(q)$ be the normal form of $Z_1(q)$, with diagonal elements

$$\frac{e_k(q)}{\psi_k(q)}$$

in lowest terms. Since $Z_1(q)$ is of rank n , none of these vanish identically.

We first claim that $\delta(Z) = \delta(Z_1)$. The poles p_0 of Z are exactly the points

$$p_0 = T(q_0)$$

where q_0 runs over the poles of Z_1 . At each pole,

$$S^*(Z, p_0) = S_T^*(Z, p_0) = S^*(Z_1, q_0)$$

by 5.35. Hence $\delta(Z, p_0) = \delta(Z_1, q_0)$ and the result follows by addition. Similarly, then, $\delta(Z^{-1}) = \delta(Z_1^{-1})$.

Next we assert that $\delta(Z_1)$ is just the degree of the polynomial

$$\psi_1(q)\psi_2(q) \cdots \psi_n(q).$$

For $\delta(Z_1, q_0)$ is the exponent of $(q - q_0)$ in this polynomial, and the zeros of this polynomial are exactly the poles of $Z_1(q)$.

We observe that if

$$W_1(q) = A(q)Z_1(q)B(q),$$

then

$$W_1^{-1}(q) = B^{-1}(q)Z_1^{-1}(q)A^{-1}(q).$$

This then is the result of polynomial operations on $Z_1^{-1}(q)$, and has diagonal elements

$$\frac{\psi_k(q)}{e_k(q)}. \quad (1)$$

Clearly by arranging these in reverse order, we have a normal form. This is 5.13. Hence the functions (1) are the diagonal elements of the normal form of $Z_1^{-1}(q)$. The argument above applied to $Z_1^{-1}(q)$ then shows that $\delta(Z_1^{-1})$ is the degree of

$$e_1(q) \cdots e_n(q).$$

Finally, we note the determinant relation

$$|W_1(q)| = |A(q)| |Z_1(q)| |B(q)| = (\text{constant}) \times |Z_1(q)|,$$

since the determinants of A and B are constant. Now $Z_1(q)$ has no pole at $q = \infty$, hence its determinant is finite there. The same is true of $Z_1^{-1}(q)$, so indeed

$$|Z_1(\infty)| = 0.$$

Now by direct calculation

$$|W_1(q)| = \frac{e_1(q) \cdots e_n(q)}{\psi_1(q) \cdots \psi_n(q)}.$$

Since this is finite and not zero at $q = \infty$, the numerator and denominator are of the same degree. Hence

$$\delta(Z) = \delta(Z_1) = \text{degree}(\Pi\psi_k) = \text{degree}(\Pi e_k) = \delta(Z_1^{-1}) = \delta(Z^{-1}).$$

VI. THE EXACT COUNT OF REACTIVE ELEMENTS

6.0 We showed in the inductive argument of 4.07 that the Brune process constructs a realization for a given $Z(p)$ which uses exactly $\delta(Z)$ reactive elements. To establish 2.18, we must still show that no network with fewer than $\delta(Z)$ reactive elements can do this. To prove this, we shall show that if $Z(p)$ is the impedance matrix of a network containing x reactive elements, then

$$\delta(Z) \leq x. \quad (1)$$

We shall, in fact, in this Section show somewhat more than (1). The demonstration of (1) requires enough calculation that is as easy to prove the following extension of 2.18.

6.01 *Theorem:* Given any linear correspondence L , (I, 6.2), which PR, (I, 16.71), there exists a number $\delta(L)$ such that

- (i) The realization process outlined in (I, 8) and 4.07 of this Part constructs with $\delta(L)$ reactive elements a network realizing a member of the Cauwer class of L .
- (ii) If L is the correspondence established by the Cauwer class of a physical network which contains x reactive elements, then

$$\delta(L) \leq x.$$

The proof is divided among the remaining paragraphs of this Section. We maintain here a strict distinction between geometric objects and their concrete coordinate representations.

6.02 We observe at once that if a $\delta(L)$ exists satisfying (i) and (ii), then it must be unique because it is exactly the minimum number of reactive elements required to realize any representative of the Cauer class L . No particular pains then need be taken as we go along to verify that the value of $\delta(L)$ arrived at is in fact independent of the mode of defining it.

6.1 Given a PR geometrical linear correspondence L between \mathbf{V} and \mathbf{K} , there is a frame which reduces L in the sense of (I, 13.02). In this frame we have the dual decomposition

$$\mathbf{V} = \mathbf{V}_{L0} \oplus \mathbf{V}_2 \oplus \mathbf{V}_1$$

$$\mathbf{K} = \mathbf{K}_1 \oplus \mathbf{K}_2 \oplus \mathbf{K}_{L0}$$

in which each subspace is real and spanned by selected basis vectors. Furthermore,

$$\mathbf{V}_L = \mathbf{V}_{L0} \oplus \mathbf{V}_2,$$

$$\mathbf{K}_L = \mathbf{K}_2 \oplus \mathbf{K}_{L0},$$

Finally, if r is the dimension of \mathbf{V}_2 and \mathbf{K}_2 , there is an $r \times r$ PR matrix $[Z_1(p)]$ such that, when

$$[v_2, k_2] \in L(p)$$

and

$$v_2 \in \mathbf{V}_2, \quad k_2 \in \mathbf{K}_2,$$

then

$$[v_2] = [Z_1(p)][k_2].$$

Here the r -tuples are those representing v_2 and k_2 as elements of \mathbf{V}_2 and \mathbf{K}_2 in the chosen frame.

6.11 *Definition:* We define $\delta(L)$ by

$$\delta(L) = \delta([Z_1]),$$

where $[Z_1(p)]$ is the matrix described above.

6.12 This number $\delta(L)$ is the number of reactive elements used when the Brune process is applied to realize $[Z_1(p)]$. (This is 4.07). Then, however, by the argument of (I, 8.5), the representative $[L]$ of L in the particular frame in question can be realized by adjoining open and short circuits to a realization of $[Z_1(p)]$. This operation adds no new reactive elements. Neither does the operation of converting $[L]$ to any

Cauer equivalent $[L]_1$ by the use of ideal transformers. Therefore the particular $\delta(L)$ we have defined—which depends for its definition upon a somewhat arbitrary choice of coordinate frame—satisfies (i) of 6.01.

6.2 Lemma: Let L be a PR geometrical linear correspondence between \mathbf{K} and \mathbf{V} , and M another between spaces \mathbf{J} and $\mathbf{U} = \mathbf{J}^*$ obtained by restricting L as in (I, 18). Then

$$\delta(M) \leq \delta(L).$$

Proof: We use the results and notation of (I, 18). In particular, C is a real constant operator from \mathbf{J} to \mathbf{K} , C^* its adjoint from \mathbf{V} to \mathbf{U} , and the pairs of $M(p)$ are those pairs

$$[u, j]$$

such that

$$u = C^*v \quad \text{and} \quad [v, Cj] \in L(p).$$

Choose a frame in \mathbf{V} and \mathbf{K} which reduces L as in 6.1. We recall that \mathbf{J}_M consists of all vectors $j \in \mathbf{J}$ such that $Cj \in \mathbf{K}_L$ (I, 18.31). Let \mathbf{J}_2 consist of all $j \in \mathbf{J}$ such that

$$Cj \in \mathbf{K}_2.$$

Let \mathbf{J}_3 consist of all $j \in \mathbf{J}$ such that

$$Cj \in \mathbf{K}_{L0}.$$

Then \mathbf{J}_2 and \mathbf{J}_3 are disjoint and both are subspaces of \mathbf{J}_M . We can therefore write

$$\mathbf{J}_M = \mathbf{J}_2 \oplus \mathbf{J}_3 \oplus \mathbf{J}_4,$$

after a suitable choice of \mathbf{J}_4 .

We now claim that

$$\mathbf{J}_3 \oplus \mathbf{J}_4 \subset \mathbf{J}_{M0}. \quad (1)$$

For we have if $j \in \mathbf{J}_M$ that, uniquely,

$$j = j_2 + j_3 + j_4,$$

where $j_i \in \mathbf{J}_i$. Therefore

$$Cj = Cj_2 + Cj_3 + Cj_4$$

where by construction $Cj_2 \in \mathbf{K}_2$, $Cj_3 \in \mathbf{K}_{L0}$, and, necessarily, then $Cj_4 = 0$. If $j_2 = 0$, therefore, $Cj \in \mathbf{K}_{L0}$ and

$$[0, Cj] \in L(p).$$

Therefore

$$[C^*0, j] = [0, j] \in M(p).$$

this proves (1).

We can now write

$$\mathbf{J}_M = \mathbf{J}_{21} \oplus \mathbf{J}_{20} \oplus \mathbf{J}_0 \quad (2)$$

where

$$\begin{aligned} \mathbf{J}_2 &= \mathbf{J}_{21} \oplus \mathbf{J}_{20}, \\ \mathbf{J}_{20} &= \mathbf{J}_2 \cap \mathbf{J}_{M0}, \\ \mathbf{J}_0 &= \mathbf{J}_3 \oplus \mathbf{J}_4, \\ \mathbf{J}_{M0} &= \mathbf{J}_{20} \oplus \mathbf{J}_0. \end{aligned} \quad (3)$$

Choosing an arbitrary \mathbf{J}_5 disjoint from \mathbf{J}_M , we can write, using (2) and (3),

$$\mathbf{J} = \mathbf{J}_5 \oplus \mathbf{J}_{21} \oplus \mathbf{J}_{M0}, \quad (4)$$

where

$$\mathbf{J}_M = \mathbf{J}_{21} \oplus \mathbf{J}_{M0}.$$

Using the arguments of (I, 12.3), we find that the decomposition of \mathbf{U} dual to (4) is, because M is PR,

$$\mathbf{U} = \mathbf{U}_{M0} \oplus \mathbf{U}_{21} \oplus \mathbf{U}_3 \quad (5)$$

where

$$\mathbf{U}_M = \mathbf{U}_{M0} \oplus \mathbf{U}_{21}.$$

As in (I, 12.3) we can now introduce a frame appropriate to the decomposition indicated in (4) and (5) and obtain a matrix $[Z_{21}(p)]$ describing the correspondence between \mathbf{J}_{21} and \mathbf{U}_{21} . Say this is an $m \times m$ matrix, m being the dimension of \mathbf{J}_{21} . We can define

$$\delta(M) = \delta([Z_{21}]).$$

Let \mathbf{J}_2 have dimension m_1 . By (3), if we border $[Z_{21}(p)]$ by $m_1 - m$ rows and columns of zeros, to obtain an $m_1 \times m_1$ matrix $[Z_2(p)]$, we can interpret $[Z_2(p)]$ as follows:

Given $j \in \mathbf{J}_2$, it can be represented by an m_1 -tuple $[j]$ in the basis in that subspace. Then the m_1 -tuple

$$[u] = [Z_2(p)][j] \quad (6)$$

represents in the dual basis in $(\mathbf{J}_2)^0$ a vector $u \in \mathbf{U}_{21}$ such that

$$[u, j] \in M(p).$$

Now this u necessarily is of the form

$$u = C^*v, \quad (7)$$

where

$$[v, Cj] \in L(p).$$

But $j \in \mathbf{J}_2$, so $Cj \in \mathbf{K}_2$, so v may be taken to be an element of \mathbf{V}_2 , with components

$$[v] = [Z_1(p)][Cj] \quad (8)$$

in the basis therein.

We have bases now in \mathbf{V} , \mathbf{K} , \mathbf{U} , and \mathbf{J} , each of which has a set of basis vectors spanning, respectively, \mathbf{V}_2 , \mathbf{K}_2 , $(\mathbf{J}_2)^0$, and \mathbf{J}_2 . By definition of \mathbf{J}_2 , and by (7) and (8), C operates from \mathbf{J}_2 to \mathbf{K}_2 , and C^* from \mathbf{V}_2 to $(\mathbf{J}_2)^0$. Hence in these respective bases C and C^* may be represented by $m_1 \times m_1$ matrices. In these bases then, from (7) and (8),

$$[u] = [C^*][v] = [C^*][Z_1(p)][C][j].$$

Comparing this with (6), we have

$$[Z_2(p)] = [C^*][Z_1(p)][C].$$

Hence by definitions and 5.26,

$$\delta(M) = \delta([Z_2]) \leq \delta([Z_1]) = \delta(L).$$

This is the assertion to be proved.

6.3 We can now turn to (ii) of 6.01. We follow the synthesis procedure of (I, 19), as modified in the remarks of 3.2.

Consider a network constructed from x reactive elements, r resistors, and some ideal transformers. As in (I, 19.2), the synthesis of this network begins by juxtaposing the $r + x$ two poles and the ideal transformers, all as separate devices. The correspondence $[L]$ established by this juxtaposition is exhibited in (I, 19.2) as one described by a diagonal matrix $[Z_d(p)]$ juxtaposed with one described by certain ideal transformers. A frame which reduces this correspondence as in 6.1 can be found by a change of basis wholly within those subspaces in which the ideal transformers operate. Hence the degree $\delta(L)$ of this correspondence in exactly $\delta([Z_d])$ which, by 5.29, is x .

Now let $[M]$ be the concrete linear correspondence established by the

network to be synthesized. Then (I, 19.3, 19.4) show that $[M]$ is obtained by two successive restrictions upon $[L]$. Hence by 6.2

$$\delta(M) \leq \delta(L) = x.$$

Q.E.D.

BIBLIOGRAPHY

1. M. Bayard, "Synthèse des Réseaux Passifs a un Nombre Quelconque de Paires de Bornes Connaissant Leurs Matrices d'Impédance ou d'Admittance," *Bulletin, Société Française des Electriciens*, **9**, 6 series, Sept. 1949.
2. O. Brune, *Jour. Math. and Phys.*, *M.I.T.*, **10**, Oct. 1931, pp. 191-235.
3. W. Cauer, *Ein Reaktanztheorem*, *Sitzungsberichte Preuss. Akad. Wissenschaft*, Heft 30/32, 1931.
4. W. Cauer, "Die Verwirklichung von Wechselstromwiderständen vorgeschriebener Frequenzabhängigkeit," *Archiv für Elektrotechnik*, **17**, 1926.
5. W. Cauer, "Ideale Transformatoren und Lineare Transformationen," *Elektrische Nachrichten-Technik*, **9**, May, 1932.
6. S. Darlington, *Journal of Mathematics and Physics*, *M.I.T.*, **18**, No. 4, Sept. pp. 257-353.
7. R. M. Foster, *Bell System Tech. J.*, April, 1924, pp. 259-267.
8. C. M. Gewertz, *Network Synthesis*, Baltimore, 1933.
9. P. R. Halmos, *Finite Dimensional Vector Spaces*, Princeton, 1942.
10. Y. Oono, "Synthesis of a Finite $2n$ -Terminal Network by a Group of Networks Each of Which Contains Only One Ohmic Resistance," *Jour. Inst. Elec. Comm. Eng. of Japan*, March, 1946. Reprinted in English in the *Jour. Math. and Phys.*, *M.I.T.*, **29**, Apr., 1950.
11. Y. Oono, "Synthesis of a Finite $2n$ -Terminal Network as the Extension of Brune's Theory of Two-Terminal Network Synthesis," *Jour. Inst. Elec. Comm. Eng. of Japan*, Aug., 1948.
12. J. L. Synge, "The Fundamental Theorem of Electrical Networks," *Quarterly of Applied Mathematics*, **9**, No. 2, July, 1951.
13. R. Bott, and R. J. Duffin, "Impedance Synthesis without the Use of Transformers," *Jour. Appl. Phys.*, **20**, Aug., 1949, p. 816.
14. H. W. Bode, *Network Analysis and Feedback Amplifier Design*, New York, 1945.
15. M. Bôcher, *Introduction to Higher Algebra*, New York, 1930.
16. C. C. MacDuffee, *An Introduction to Abstract Algebra*, New York, 1940.

Abstracts of Bell System Technical Papers*

Not Published in This Journal

The Effect of Inhomogeneities on the Electrical Properties of Diamond.

A. J. AHEARN¹. *Phys. Rev.*, **84**, pp. 798–802, Nov. 15, 1951.

To account for the non-uniformities in the electrical properties of diamond, particularly those observed in bombardment conduction, the proposal is made that the well-known lattice imperfections are not distributed homogeneously in the physical crystal, and that the resulting fluctuations in the height of the energy bands relative to the Fermi level might produce interspersed “pools of mobile charge” separated by barriers within the diamond. These pools and barriers should lead to dielectric losses at high frequencies. A single conducting channel, in series with a barrier, could be represented by a series resistance R_s and capacity C_s or by the equivalent parallel resistance R_p and capacity C_p .

With some, but not all, diamonds measurable dielectric losses at 70 mc/sec were observed. R_p varied from 5×10^6 ohms, the limit of measurement, to 4×10^5 ohms. Furthermore, the proposed model suggests that, in some cases, these barriers might be sufficiently lowered to establish a dc conducting channel all the way through a crystal. With a few of the lossy diamonds precisely this phenomenon of “high conduction” has been observed, in which a resistance of the order of a megohm is obtained with a dc voltage applied. This current appears abruptly in time but it lags behind the application of the voltage. This lag is influenced by irradiation with light or alpha-particles or by previous treatment.

The proposed pools of mobile charge are a sufficient but not necessary description of the dielectric loss observations, but the high conduction phenomenon lends further support to this idea of conducting channels with barriers in lossy diamonds. Such localized conducting channels would introduce inhomogeneities into an otherwise uniform electric field applied across an insulator. In bombardment conduction, measurements of counting efficiency could be very sensitive to field inhomogeneities.

Under alpha-particle bombardment a large variation in counting efficiency over the surface of typical diamonds is shown. In a group of 20 diamonds, most of those that exhibited definite losses also had high (≥ 25 per cent) counting efficiencies in some region, and the majority of the remainder had low counting efficiencies. These experiments lend further support to the suggestion that in-

* Certain of these papers are available as Bell System Monographs and may be obtained on request to the Publication Department, Bell Telephone Laboratories, Inc., 463 West Street, New York 14, N. Y. For papers available in this form, the monograph number is given in parentheses following the date of publication, and this number should be given in all requests.

¹ Bell Telephone Laboratories.

homogeneous fields at least partially account for the inhomogeneities in bombardment conduction.

Serious errors in the normal estimates of range and mobility of electrons or holes in insulators can be introduced by neglecting these field inhomogeneities.

Diffusion in Alloys and the Kirkendall Effect. J. BARDEEN¹ and C. HERRING¹. Pp. 87-111. *Am. Soc. for Metals*. Atom movements; a seminar . . . held during the thirty-second National Metal Congress and Exposition, Chicago, Oct. 21-27, 1950. Cleveland, Ohio, Am. Soc. for Metals, 1951. 240 p.

Some Roots of an Equation Involving Bessel Functions. B. P. BOGERT¹. *Jl. Math. Phys.*, **30**, pp. 102-105, July, 1951. (Monograph 1903).

Creep Test Methods for Determining Cracking Sensitivity of Polyethylene Polymers. W. C. ELLIS¹ and J. D. CUMMINGS¹. *A.S.T.M. Bull.*, No. 178, pp. 47-49, Dec., 1951.

Conventional creep testing methods for evaluating the cracking sensitivity of polyethylene polymers are described. The tests show that sensitivity to cracking in the presence of an active agent decreases with increasing average molecular weight of the polymer. For a given stress condition and environment, there appears to be a threshold value of stress and strain for the occurrence of cracking.

Observer Reaction to Video Crosstalk. A. D. FOWLER¹. *J. Soc. Motion Picture and Television Engrs.*, **57**, pp. 416-424, Nov., 1951. (Monograph 1928).

Presented here are results of tests to determine how much video crosstalk can be tolerated in black-and-white television pictures. Experienced observers viewed a television picture and rated the disturbing effects of controlled amounts of crosstalk from another video system. Crosstalk coupling was simulated by a network which permitted changes in frequency characteristic as well as in coupling loss. Tolerable limits for crosstalk coupling are derived from the test results.

Mass Spectrometric Studies of Molecular Ions in the Noble Gases. J. A. HORNBECK¹ and J. P. MOLNAR¹. *Phys. Rev.*, **84**, pp. 621-625, Nov. 15, 1951.

Molecular ions of the rare gases (He_2^+ , Ne_2^+ , Ar_2^+ , Kr_2^+ , and Xe_2^+) produced by electron impact at gas pressures from 10^{-4} to 10^{-2} mm Hg have been studied with a small mass spectrometer. The ion intensity increased linearly with electron current and with the square of the gas pressure. The form of the ionization versus electron energy curves resembles closely curves of excitation probability by electron collision. The appearance potentials of the molecular ions were less

¹ Bell Telephone Laboratories

than those of the atomic ions by $1.4^{+0.7}_{-0.2}$ volts in He, $0.7^{+0.7}_{-0.3}$ volt in Ne, $0.7^{+0.7}_{-0.2}$ volt in A, $0.7^{+0.7}_{-0.3}$ volt in Kr. These results can be interpreted, we believe, only by assuming that the process of formation of the molecular ions observed in this experiment is, using helium as an example, an excitation by electron impact, $\text{He} + e + \text{K.E.} \rightarrow \text{He}^* + e$, followed by the collision process, $\text{He}^* + \text{He} \rightarrow \text{He}_2^+ + e$, where He^* stands for a helium atom raised to a high-lying excited state. Our results differ from those of Arnot and M'Ewen on helium particularly in that they reported the appearance potential low enough to permit metastable atoms to form molecular ions.

The Drift Velocities of Molecular and Atomic Ions in Helium, Neon, and Argon. J. A. HORNBECK¹. *Phys. Rev.*, **84**, pp. 615-620, Nov. 15, 1951.

Drift velocity measurements as a function of E/p_0 , the ratio of field strength to normalized gas pressure, are presented for atomic and molecular ions of He, Ne, and A in their respective parent gases. Identification of the molecular ions is based upon the time resolution of the apparatus and the dependence of ion concentration on pressure, applied voltage, and gas purity. Extrapolation of the low field measurements to zero field yields mobility values for atomic ions, $\mu_0 (\text{He}^+) = 10.8 \text{ cm}^2/\text{volt sec}$, $\mu_0 (\text{Ne}^+) = 4.4$, and $\mu_0 (\text{A}^+) = 1.63$ in good agreement with theory: Massey and Mohr compute $\mu_0 (\text{He}^+) = 11$, and Holstein gives $\mu_0 (\text{Ne}^+) = 4.1$ and $\mu_0 (\text{A}^+) = 1.64$. Drift velocity data at low field for the molecular ions agree within experimental error with data of Tyndall and Powell (He), and Munson and Tyndall (Ne and A), which they assigned to atomic ions. A qualitative description in terms of ion-atom interaction forces is given for the observed field variation of the atomic ion drift velocities up to high E/p_0 .

Checking Analogue Computer Solutions. E. LAKATOS¹. *Proc. Inst. Radio Engrs.*, **39**, p. 1571, Dec., 1951.

Experimental Heat Contents of SrO , BaO , CaO , BaCO_3 and SrCO_3 at High Temperatures. Dissociation Pressures of BaCO_3 and SrCO_3 . J. J. LANDER¹. *J. Am. Chem. Soc.*, **73**, pp. 5794-5797, Dec., 1951. (Monograph 1930).

The high temperature heat contents of SrO , BaO , CaO , BaCO_3 and SrCO_3 have been measured using the "drop" method. Values have been obtained for the heats of the transitions of the carbonates. The dissociation pressures of the carbonates have been measured to pressures below 0.1 mm and values calculated for lower pressures from the observed heat contents and observed dissociation pressures at higher temperatures.

Electron-Hole Production in Germanium by Alpha-Particles. K. G. MCKAY¹. *Phys. Rev.*, **84**, pp. 829-832, Nov. 15, 1951.

The number of electron-hole pairs produced in germanium by alpha-particle

¹ Bell Telephone Laboratories

bombardment has been determined by collecting the internally produced carriers across a reverse-biased n - p junction. No evidence is found for trapping of carriers in the barrier region. Studies of individual pulses show that the carriers are swept across the barrier in a time of less than 2×10^{-8} sec. The counting efficiency is 100 per cent. The energy lost by an alpha-particle per internally produced electron-hole pair is 3.0 ± 0.4 ev. The difference between this and the energy gap is attributed to losses to the lattice by the internal carriers. It is concluded that recombination due to columnar ionization is negligible in germanium.

The n-p-n Junction as a Model for Secondary Photoconductivity. K. G. MCKAY¹. *Phys. Rev.*, **84**, pp. 833-835, Nov. 15, 1951.

A germanium n - p - n junction with the p region floating, has been subjected to alpha-particle bombardment. The transient currents resulting from individual incident alphas have been studied. This enables one to study the rate of decay of excess holes in the p -region. This decay time appears to increase with applied bias, pass through a maximum, and eventually approach a constant value. The total charge flowing across the unit, as a result of the bombardment by a single alpha-particle, may become large; quantum yields of greater than 60 have been observed. The unit possesses many of the important characteristics of materials which exhibit "secondary photoconductivity." It is concluded that various forms of n - p - n barriers must therefore play an important role in such materials and that their understanding can be greatly facilitated by studies of n - p - n barriers in germanium.

Frequency Detection and Speech Formants. E. PETERSON¹. *Acoustical Soc. Am., Jl.*, **23**, pp. 668-674, Nov., 1951.

This study is aimed primarily at evaluating the utility of axis-crossing detectors in tracking speech formants. Detectors of the usual type are found subject to an error, fundamental in nature. To remove this source of error speech is modulated up in frequency as a single sideband before limiting and detecting processes are applied. Experimental results with this carrier type of detector on a small number of speech samples are presented, and compared with spectrograms. Conclusions are that the average axis-crossing rates cannot be trusted in general to follow specific formants, whether the speech is normal or differentiated. But when the formants are sufficiently localized by frequency selectivity, prospects of tracking the lower formants look promising.

Transistor Circuit Design. G. RAISBECK¹. *Electronics*, **24**, pp. 128-132, 134, Dec., 1951. (Monograph 1932).

How to derive amplifier, oscillator, modulator and multi-vibrator transistor circuits from known vacuum-tube circuits. Technique, known as duality, is explained in detail and may be applied to any complex vacuum-tube circuit to find the corresponding transistor circuit.

Communication Theory—Exposition of Fundamentals. C. E. SHANNON¹. Pp. 44-47. General treatment of the problem of Coding. Pp. 102-104.

¹ Bell Telephone Laboratories

Great Britain. Ministry of Supply. Symposium on Information Theory. Report of Proceedings held . . . Royal Soc., Burlington House, Lond., Sept. 26-29, 1950.

On the Relation Between the Sound Fields Radiated and Diffracted by Plane Obstacles. F. M. WIENER¹. *J. Acoust. Soc. Am.*, **23**, pp. 697-700, Nov., 1951.

In the past, acoustic diffraction and radiation problems have often been treated separately, although their intimate connection is clear from theory. In the case of plane piston radiators and plane rigid scatterers exposed to a perpendicularly incident plane wave, this connection becomes particularly simple and useful. It is easy to show that the radiated sound field is everywhere the same as the field scattered (diffracted) in the diffraction case, except for a factor of proportionality. It is also shown that the reaction of the medium on the radiator, as expressed by the mechanical radiation impedance, is equal to the force per unit incident pressure exerted on the same obstacle, held rigid as a scatterer, except for a factor of proportionality. By way of illustration, the foregoing principles are applied to the important case of the circular disk.

Magnetic Modulators. E. P. FELCH¹, V. E. LEGG¹, and F. G. MERRILL¹. References. *Electronics*, **25**, pp. 113-117, Feb., 1952.

Conversion of low-level, low-frequency or dc signals to ac signals capable of being amplified by conventional means is accomplished by magnetic-amplifier-type device that combines high efficiency and reliability with extreme ruggedness.

Conservation of Nickel. G. R. GOHN¹. *A.S.T.M., Bull.*, No. 179, p. 32, Jan., 1952.

The Mechanism of Electrolytic Rectification. H. E. HARING¹. *Electrochem. Soc., Jl.*, **99**, pp. 30-37, Jan., 1952. (Monograph 1929).

An electrochemical theory is proposed for rectification, as exemplified by the tantalum (or aluminum) electrolytic rectifier and capacitor. A detailed consideration of the mechanism of formation of the oxide film which constitutes the rectification barrier leads to the conclusion that this barrier consists of an electrolytic polarization, in the form of a concentration gradient of excess metal ions, permanently fixed or "frozen" in position in an otherwise insulating matrix of electrolytically-formed oxide. The physical structure which has been described functions as (a) a current-blocking ionic space charge or (b) a current-passing electronic semiconductor, depending solely upon the direction of the applied voltage. The movement of electrons only is required. An explanation for breakdown of the barrier at excessively high voltages is suggested. This explanation may be applicable to dielectric breakdown of other kinds.

¹ Bell Telephone Laboratories

Nullification of Space-Charge Effects in a Converging Electron Beam by a Magnetic Field. M. E. HINES¹. *Proc. Inst. Radio Engrs.*, **40**, pp. 61-64, Jan., 1952. (Monograph 1935).

This paper presents the conditions necessary for maintaining a uniformly converging conical electron beam in the presence of space charge. It is an extension of the Brillouin focusing condition to conical flow, requiring a converging rather than a uniform magnetic field. In this type of electron flow, the diverging effects of space charge are balanced against magnetic reaction forces for reasonably small cone angles of convergence. Though the balance of forces is exact only for infinitesimal angles, it is reasonably accurate for cones of half-angle as great as 10 degrees. The minimum beam size will be limited only by the effects of thermal velocities, by gun aberrations, and by the magnetic field obtainable.

Continuous Motion Picture Projector for Use in Television Film Scanning. A. G. JENSEN¹, R. E. GRAHAM¹, and C. F. MATTKE¹. Bibliography. *J. Soc. Motion Picture and Television Engrs.*, **58**, pp. 1-21, Jan., 1952.

The projector used for this experiment drives a 35-mm motion picture film at the standard (nonintermittent) speed of 24 frame/sec and produces a television signal of 525 lines and 30 frames interlaced 2 to 1. The projector utilizes a system of movable plane mirrors mounted on a rotating drum and controlled by a single stationary cam. Vertical jitter in the television image is minimized by means of an electronic servo system operating on the film sprocket holes, resulting in a residual vertical motion of about 1/2000 of a picture height. A second electronic servo system is incorporated to suppress flicker. The combination of this scanner and a high-grade monitor is capable of producing a television picture with a resolution corresponding to about 8 mc and with good tone rendition over a range up to 200 to 1.

Low Temperature Polymorphic Transformation in WO₃. B. T. MATTHIAS¹ and E. A. WOOD¹. *Phys. Rev.*, **84**, p. 1255, Dec. 15, 1951.

The Concentration of Molecules on Internal Surfaces in Ice. E. J. MURPHY¹. *J. Chem. Phys.*, **19**, pp. 1516-1518, Dec., 1951.

In this paper the experimental expression for the "local conductivity" of ice is given. This expression has two terms, one of which has already been discussed and brought into close relation with the structure of ice, that is, with its heat of sublimation and its lattice constant. This paper brings out another relation, deriving it from the second term of the experimental expression. It is concluded from an analysis outlined here that the second term of the local conductivity gives the concentration of molecules in "internal surfaces". For the specimen of ice to which this method was applied the concentration of molecules on internal surfaces comes out as 1.03×10^{17} molecules/cc. This is proposed as a new method of studying imperfections (internal surfaces) in dielectric crystals, and one which seems to be well suited to this purpose. It gains its advantages from

¹ Bell Telephone Laboratories

the fact that it is not dependent upon the regularity of the imperfections, as in x-ray diffraction methods, or upon the connectivity of the system of internal surfaces, as in direct current conduction.

Meditations on Physics Today. J. R. PIERCE¹. *Phy. Today*, **5**, p. 3, Jan., 1952.

Stabilization of Dielectrics Operating Under Direct Current Potential. H. A. SAUER¹, D. A. McLEAN¹, and L. EGERTON¹. *Ind. Eng. Chem.*, **44**, pp. 135-140, Jan., 1952.

¹ Bell Telephone Laboratories.

Contributors to this Issue

E. N. GILBERT, B.S., Queens College, 1943; Ph.D., Massachusetts Institute of Technology, 1948. M.I.T. Radiation Laboratory, 1944-46. Bell Telephone Laboratories, 1948-. Dr. Gilbert's first assignment was in a group studying information theory, and in 1949 he joined a group concerned with switching theory. Member of the American Mathematical Society.

I. L. HOPKINS, B.S., Massachusetts Institute of Technology, 1927; Bell Telephone Laboratories, 1927-. For eighteen years, Mr. Hopkins designed testing equipment and tested insulating materials. Right after World War II, he tested and developed special-purpose rubber compounds, and since 1948 he has been conducting research in the physical properties of polymers.

W. A. MALTHANER, B.E.E., Rensselaer Polytechnic Institute, 1937. Bell Telephone Laboratories, 1937-. Mr. Malthaner is currently engaged in research on new automatic telephone central-office systems, inter-office signaling systems, and subscriber dialing and supervisory arrangements. Until World War II, when he worked on the development of automatic fire control systems and fire control radar, Mr. Malthaner tested and developed central office circuits and switching systems. Associate of the American Institute of Electrical Engineers. Member of the Institute of Radio Engineers, Tau Beta Pi and Sigma Xi.

WARREN P. MASON, B.S. in E.E., University of Kansas, 1921; M.A., Ph.D., Columbia, 1928. Bell Telephone Laboratories, 1921-. Dr. Mason has been engaged in investigating the properties and applications of piezoelectric crystals, in the study of ultrasonics, and in mechanics. Fellow of the American Physical Society, Acoustical Society of America and Institute of Radio Engineers and member of Sigma Xi and Tau Beta Pi.

BROCKWAY McMILLAN, B.S., Massachusetts Institute of Technology, 1936; Ph.D., Massachusetts Institute of Technology, 1939; Instructor of Mathematics, Massachusetts Institute of Technology, 1936-39; Procter

Fellow and Henry B. Fine Instructor in Mathematics, Princeton University, 1939-42; U.S.N.R., 1942-46, studying exterior ballistics of guns and rockets; Los Alamos Laboratory, spring 1946; Bell Telephone Laboratories, 1946-. Dr. McMillan has been engaged in mathematical research and consultation work. Member of American Mathematical Society, Institute of Mathematical Statistics, and A.A.A.S.

JAMES Z. MENARD, B.S., Arkansas State Teachers College, 1941. U. S. Army, 1941-46. Bell Telephone Laboratories, 1946-. Mr. Menard has been engaged in the development of magnetic recording equipment and audio equipment for telephone plant applications.

J. A. MORTON, B.S. in E.E., Wayne University, 1935; M.S., University of Michigan, 1936. Bell Telephone Laboratories, 1936-. Mr. Morton is currently in charge of the development of the transistor and other semi-conductor devices. In the past he has been concerned with research on coaxial cables, microwave amplifier circuits, radar receivers, and with vacuum tube development. He designed a microwave tube used in the New York-San Francisco microwave relay system. Member of the I.R.E., Eta Kappa Nu, Alpha Delta Psi, Mackenzie Honor Society, Phi Kappa Phi, and Sigma Xi.

H. EARLE VAUGHAN, B.S. in C.E., Cooper Union, 1933. Bell Telephone Laboratories, 1928-. Since World War II, Mr. Vaughan has been investigating switching systems and high speed signaling means. In the past he studied voice operated devices and fundamental effects of speech and noise on voice-frequency signaling systems. During World War II, he was engaged in government projects, conducting research on anti-aircraft computers and fire control radars.

SAMUEL D. WHITE, B.S. in E.E., Rutgers University, 1927; E.E., Rutgers University, 1932; Bell Telephone Laboratories, 1927-. Until 1939, Mr. White was a member of the acoustical research department. He then entered the switching apparatus development group and is currently studying some aspects of relay problems. Member of I.R.E., Acoustical Society of America, and Sigma Xi.

H E B E L L S Y S T E M

Technical Journal

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

VOLUME XXXI

JULY 1952

NUMBER 4

KANSAS CITY, MO.
PUBLIC LIBRARY
AUG - 5 1952

Thirtieth Anniversary 611

Lee de Forest and William Shockley Discuss Electronics 612

Network Synthesis Using Tchebycheff Polynomial Series
SIDNEY DARLINGTON 613

A Carrier Telegraph System for Short-Haul Applications
J. L. HYSKO, W. T. REA AND L. C. ROBERTS 666

The Type-O Carrier System
PAUL G. EDWARDS AND L. R. MONTFORT 688

Efficient Coding B. M. OLIVER 724

Statistics of Television Signals E. R. KRETZMER 751

Experiments with Linear Prediction in Television C. W. HARRISON 764

Generalized Telegraphist's Equations for Waveguides
S. A. SCHELKUNOFF 784

Photoelectric Properties of Ionically Bombarded Silicon
EDWIN F. KINGSBURY AND RUSSELL S. OHL 802

Abstracts of Bell System Papers Not Published in this Journal 816

Contributors to this Issue 820

THE BELL SYSTEM TECHNICAL JOURNAL

PUBLISHED SIX TIMES A YEAR BY THE
AMERICAN TELEPHONE AND TELEGRAPH COMPANY

195 BROADWAY, NEW YORK 7, N. Y.

CLEO F. CRAIG, *President*

CARROLL O. BICKELHAUPT, *Secretary*

DONALD R. BELCHER, *Treasurer*

EDITORIAL BOARD

F. R. KAPPEL

O. E. BUCKLEY

H. S. OSBORNE

M. J. KELLY

J. J. PILLIOD

A. B. CLARK

R. BOWN

D. A. QUARLES

F. J. FEELY

PHILIP C. JONES, *Editor*

M. E. STRIEBY, *Managing Editor*

SUBSCRIPTIONS

Subscriptions are accepted at \$3.00 per year. Single copies are 75 cents each.

The foreign postage is 65 cents per year or 11 cents per copy.

The Bell System Technical Journal

Volume XXXI

July 1952

Number 4

COPYRIGHT 1952, AMERICAN TELEPHONE AND TELEGRAPH COMPANY

Thirtieth Anniversary

Thirty years ago this month THE BELL SYSTEM TECHNICAL JOURNAL began publication. Suggested by Dr. George A. Campbell, it had been under discussion for some years. Dr. R. W. King, who had been one of its most active advocates, became its editor when the staff of the JOURNAL was established. Except for a six-year period following 1928, while he was in England, Dr. King continued as editor until he retired in 1949.

By July, 1922, when No. 1, Vol. 1 of the JOURNAL appeared, research and development was a long established practice in the Bell System. The high-vacuum electronic tube, which had already begun to revolutionize electrical communication, was itself a product of Bell System research. Since electrical communication was a still comparatively new field of study, however, its publications were widely scattered. There seemed a need for a magazine that would serve the communication engineers exclusively, and it was largely to meet this need that THE BELL SYSTEM TECHNICAL JOURNAL was launched.

In the thirty years since that time, the art and science of communication has advanced and ramified beyond anything likely to have been then foreseen. A very substantial part of this increase has originated within the Bell System, and this progress has been reflected in the pages of the TECHNICAL JOURNAL. There seems little reason to doubt that the next three decades will witness an advance at least comparable with that of the past three, and it is planned to have the JOURNAL present the work of the coming years, with perhaps even greater effectiveness than in the past. Abstracts or titles of all Bell System technical and scientific papers appearing in other publications are listed in the JOURNAL and reprints of many of these papers are available and may be obtained by subscribers. In one way or another, therefore, JOURNAL readers have access to essentially all the technical papers published by the Bell System. With this increased coverage, it is hoped that the JOURNAL will prove increasingly useful to a growing circle of readers.



Lee de Forest and William Shockley Discuss Electronics

Dr. de Forest is the inventor of the Audion from which the modern vacuum tube in its many forms and types has sprung. Dr. Shockley is the leader of the research group at Bell Telephone Laboratories whose members invented the transistor. Standing side by side these two men seem to epitomize the basic change in the pattern of our technical life which has taken place during the first half of the present century—the change from the struggling individual inventor to the great industrial scientific laboratory as the source of much of our technological advance.

Network Synthesis Using Tchebycheff Polynomial Series†

By SIDNEY DARLINGTON

(Manuscript received April 17, 1952)

A general method is developed for finding functions of frequency which approximate assigned gain or phase characteristics, within the special class of functions which can be realized exactly as the gain or phase of finite networks of linear lumped elements. The method is based upon manipulations of two Tchebycheff polynomial series, one of which represents the assigned characteristic, and the other the approximating network function. The wide range of applicability is illustrated with a number of examples.

1. INTRODUCTION

Network synthesis is the opposite of network analysis—namely, the design of a network to have assigned characteristics, as opposed to the evaluation of the characteristics of an assigned network. In general, there are specifications on the internal constitution of the network, as well as requirements relating to its external performance. A common form of the general problem is the design of a finite network of linear lumped elements, to produce an assigned gain or phase characteristic over a prescribed interval of useful frequencies. The present paper relates to this particular form.

In general, the restrictions on the network are such that the assigned performance cannot be matched exactly. This gives rise to an approximation or interpolation problem. For present purposes, the problem is: to choose a function of frequency which matches the assigned gain or phase to a satisfactory accuracy, from that special class of functions which can be realized exactly with physical finite networks of linear lumped elements. The function of frequency may be defined in terms of network singularities (natural modes and infinite loss points). The

† Presented orally, in briefer form, at the 1951 Western Convention of the Institute of Radio Engineers, and at the Symposium on *Modern Network Synthesis* sponsored by the Polytechnic Institute of Brooklyn and The Office of Naval Research, New York City, April, 1952.

interpolation problem may then be regarded as solved when a suitable set of network singularities has been obtained; for quite different techniques are used to design actual networks with these singularities.

The interpolation problem may be attacked in a number of different ways; and a variety of different techniques are, in fact, needed to cover the wide diversity of practical applications. The present topic is a fairly general way of attacking the problem, based upon manipulations of two series of Tchebycheff polynomials. The two series represent expansions of two functions of frequency—one, the ideal assigned gain or phase, the other, the network approximation to the ideal. The interpolation problem may be solved in this way because it is feasible, as will be shown, to determine network singularities from arbitrarily assigned values of coefficients in the corresponding Tchebycheff polynomial series.

The techniques to be described were derived originally from studies of the so-called potential analogy; but they can now be developed most easily without reference thereto.† In a sense they may be regarded as extensions of familiar filter theory, using Tchebycheff polynomials, to more general gain and phase functions. The extensions, however, depend upon a number of new principles. Extensions of the filter theory applied to more general problems have been noted in published papers; but those noted have not used the specific general approach employed here.‡ The wide applicability of this general approach will be illustrated by specific examples.

2. NETWORK AND TRANSMISSION FUNCTION

It will be sufficient for our present purposes to limit the discussion to the general 4-pole shown in Fig. 1. The 4-pole may be active or passive, but it must be a stable finite network of linear lumped elements. E and V are steady state ac voltages, E the driving voltage and V the response. The gain α and phase β are here defined as the real and imaginary parts of $\log V/E$.

For a finite network of lumped elements, $\alpha + i\beta$ always has the following form:

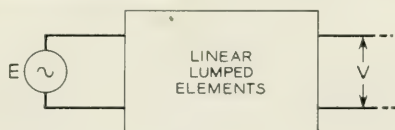
$$\alpha + i\beta = \log K \frac{(p - p'_1)(p - p'_2) \cdots}{(p - p''_1)(p - p''_2) \cdots} \quad (1)$$

† Tchebycheff polynomials are related to potential analogue charges on ellipses, as described in the author's paper "The Potential Analogue Method of Network Synthesis", Section 15.

‡ For the most part, they have used the potential analogy, in such a way that Tchebycheff polynomials do not appear at all in general applications. For examples, see methods of Matthaei², Bashkow³, and Kuh⁴.

The "frequency variable" p represents, of course, $i\omega$. The zeros p'_σ of the rational fraction are those values of p at which there is infinite loss. The poles p''_σ are the so-called natural modes, or values of p at which response V can exist in the absence of driving voltage E . The scale factor K determines the average level of transmission. The zeros, poles, and scale factor together determine the gain and phase completely.

For a physical stable network, the zeros and poles must meet certain well known restrictions, which are commonly stated in terms of locations in the complex plane for frequency variable p . Within these restrictions, the zeros and poles can be subject to arbitrary choice, say for purposes of network synthesis.



$$\alpha + i\beta = \log V/E$$

Fig. 1—A general 4-pole.

The symmetries required by the physical restrictions permit α and β to be represented separately as follows:†

$$\begin{aligned} 2\alpha &= \log K^2 \frac{(p_1'^2 - p^2)(p_2'^2 - p^2) \cdots}{(p_1''^2 - p^2)(p_2''^2 - p^2) \cdots} \\ i2\beta &= \log \frac{(p_1' - p) \cdots (p_1'' + p) \cdots}{(p_1' + p) \cdots (p_1'' - p) \cdots} \end{aligned} \quad (2)$$

These expressions hold at all real frequencies, but only at real frequencies.

3. TCHEBYCHEFF POLYNOMIALS

It is functions of these special types which we are to synthesize with the help of Tchebycheff polynomials. More generally, Tchebycheff polynomials come in various forms, and may be analyzed in various ways. For our special purposes, however, they take somewhat special forms (a little different from textbook definitions); and they are best analyzed in quite special ways.‡ It will be simplest to start with arbitrary definitions, to be justified later on by demonstrations of usefulness.

† The phase equation omits a possible 180° phase reversal, which is trivial for present purposes.

‡ For discussions of Tchebycheff polynomials from other viewpoints, see Courant and Hilbert⁵, and also a paper by Lanczos⁶ on trigonometric interpolation.

Actually, the definitions must vary with the nature of the useful frequency interval. For the present, however, it will be assumed that the useful interval extends from $\omega = 0$ to ω_c ; or more precisely, from $\omega = -\omega_c$ to $+\omega_c$ (in accordance with the symmetries of gain and phase functions). Useful intervals which do not include $\omega = 0$ require changes in the definitions, which will be taken up in Section 28.

For our present purposes, Tchebycheff polynomials T_k may be defined as follows:

$$\begin{aligned} p &= i\omega = i\omega_c \sin \phi \\ T_k &= \cos k\phi, \quad k \text{ even} \\ T_k &= i \sin k\phi, \quad k \text{ odd} \end{aligned} \quad (3)$$

The first equation defines an auxiliary angle variable, ϕ , in terms of which T_k is especially simple. The imaginary scale factor i , associated with polynomials of odd order, simplifies later analysis. In addition, it

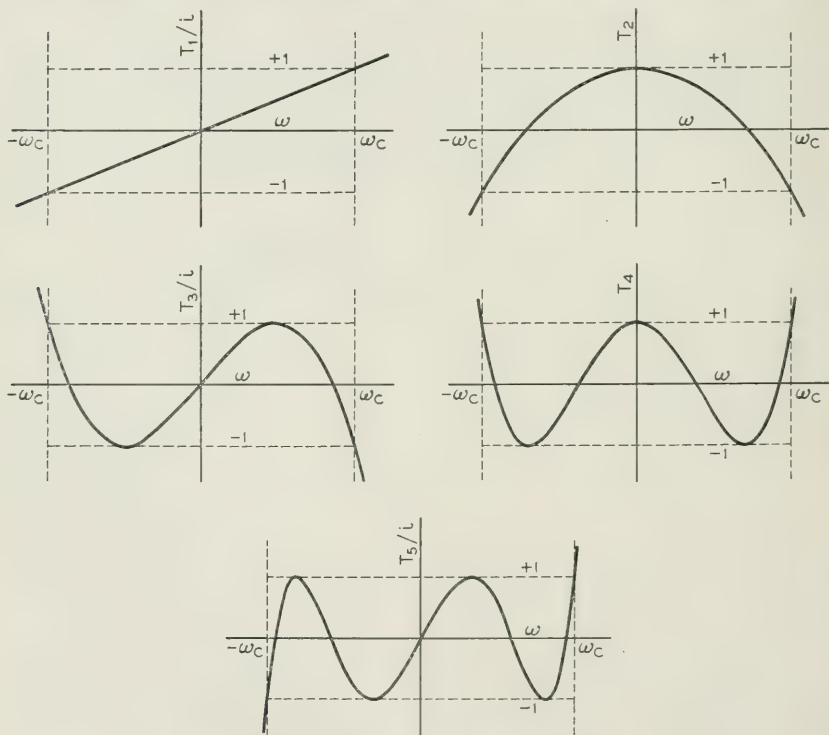


Fig. 2—Tchebycheff polynomials.

is especially appropriate for general network applications, because the odd ordered polynomials contribute to the imaginary parts of complex network functions—such as $i\beta$ in $\alpha + ib$.†

It is apparent from (3) that the Tchebycheff polynomials become simply Fourier harmonics, if they are plotted against a *distorted* frequency scale—that is, against ϕ . This means that they must be orthogonal, over that particular range of frequencies which corresponds to real values of ϕ . From the relation between ϕ and ω , it is clear that real values of ϕ cover the frequency interval between $-\omega_c$ and $+\omega_c$, which is our useful interval. In other words, the interval of orthogonality coincides with the useful frequency interval. The corresponding interval of p is of course $p = -i\omega_c$ to $+i\omega_c$.

If a given function is plotted against ϕ , instead of ω , it may be expanded in a Fourier series. Each term in the series may be replaced by a Tchebycheff polynomial, to obtain an expansion of a given function in terms of polynomials, for the specific useful interval $\omega = -\omega_c$ to $+\omega_c$. Established techniques are available for expanding experimental, or other numerical data, in Fourier series, as well as actual analytic functions.

In Fig. 2, some of the Tchebycheff polynomials are plotted against ω . The frequencies $-\omega_c$ and $+\omega_c$ are also indicated. Frequencies between these limits correspond to real values of the angle variable ϕ . If this part of the frequency scale is stretched, in the proper non-uniform way, the various “loops” not only have the same maximum values, but also the same shapes. In other words, they become periodic. More specifically, a stretch which changes the frequency scale into a ϕ scale changes the plots into $\sin k\phi$ or $\cos k\phi$.

4. TRANSFORMATION OF VARIABLE

An alternate to (3) may be obtained by relating a new variable, z , to ϕ by

$$z = e^{i\phi} \quad (4)$$

Substituting z in the exponential equivalent of $\sin \phi$, in the first equation of (3), gives an alternative definition of z directly in terms of p , namely:

$$p = \frac{\omega_c}{2} \left(z - \frac{1}{z} \right) \quad (5)$$

† A small change in the definition of ϕ would bring the definitions closer to convention, by replacing both sines by cosines (without altering T_k as a function of p). This however, would complicate our later analysis.

Substitutions in the exponential equivalents of the other sine and cosine in (3) give:

$$\begin{aligned} T_k &= \frac{1}{2} \left(z^k + \frac{1}{z^k} \right), & k \text{ even} \\ T_k &= \frac{1}{2} \left(z^k - \frac{1}{z^k} \right), & k \text{ odd} \end{aligned} \quad (6)$$

Network applications depend upon the nature of the relationship between the variable p , and the variable z . The relationship is illustrated in Fig. 3, which indicates corresponding contours in the p and z planes.

Since angle ϕ is real in the useful interval, z , as defined by (4), has unit magnitude. In equivalent conformal mapping terms, the unit circle in the z plane maps onto a segment of the axis of real frequencies in the p plane—namely the segment extending from $p = -i\omega_c$ to $+i\omega_c$. Hereafter, we shall say merely that the useful interval in the z plane is the unit circle, or $|z| = 1$. The real frequency intervals outside the useful interval map onto the imaginary axis in the z -plane.

z -plane circles with radii other than unity map onto p -plane ellipses, all with foci at $p = \pm i\omega_c$. This is reminiscent of filter theory using Tchebycheff polynomials, and in fact such a filter may be obtained by spacing z -plane mappings of natural modes uniformly around such a circle.[†]

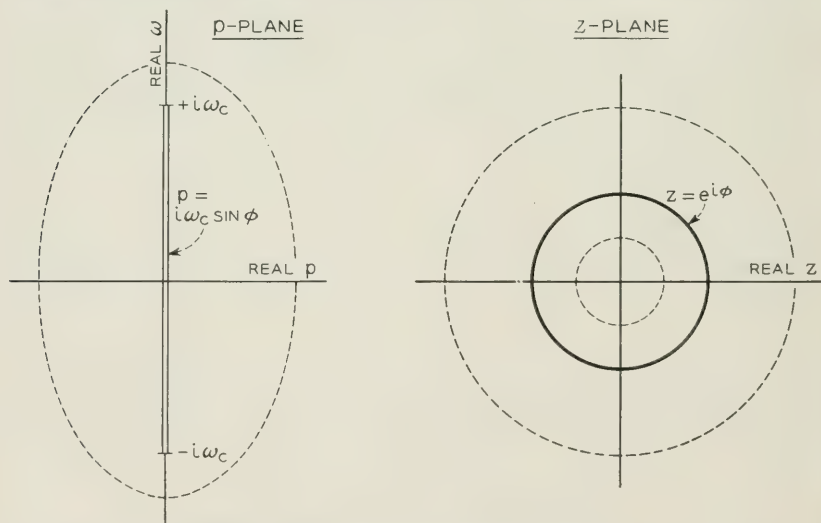


Fig. 3—The complex planes for p and z .

[†] The filter theory is developed in detail in a monograph by Wheeler⁷, which also includes an extensive bibliography.

5. Z-PLANE MAPPINGS OF NETWORK SINGULARITIES

z-plane mappings of network singularities are also an essential part of synthesis applications. The mapping z_σ of a typical zero or pole p_σ is illustrated in Fig. 4. From (5), the analytic relation must be:

$$p_\sigma = \frac{\omega_c}{2} \left(z_\sigma - \frac{1}{z_\sigma} \right) \quad (7)$$

By its quadratic nature, there must be exactly *two* values of z_σ , corresponding to one p_σ . The relation is such that replacing z_σ by $-1/z_\sigma$ leaves p_σ unchanged; and hence the two values of z_σ must be negative reciprocals, each of the other. Thus, one mapping of p_σ falls outside the unit z-plane circle, and the other inside.

A unique definition of z_σ may be obtained by requiring that z_σ must be the mapping *outside* the unit circle. Then $|z_\sigma| > 1$ by definition, and the complete definition of z_σ may be:

$$p_\sigma = \frac{\omega_c}{2} \left(z_\sigma - \frac{1}{z_\sigma} \right) \quad (8)$$

$$|z_\sigma| > 1$$

This definition is unique provided network singularities p_σ are excluded from that very special line segment of the real frequency axis which corresponds to the useful frequency interval, $-\omega_c < \omega < +\omega_c$ (where $|z_\sigma|$ would be exactly 1).

We are going to solve the interpolation problem by choosing the z_σ first, instead of the p -plane singularities p_σ , after formulating the interpolation problem in suitable z-plane terms. For this, however, we must

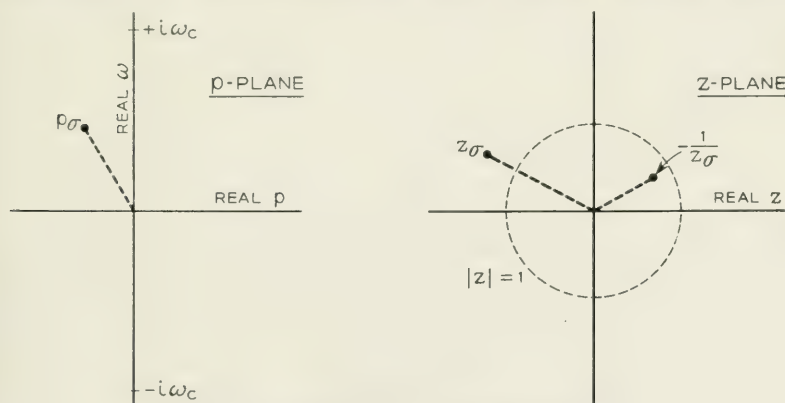


Fig. 4—Mappings of a network singularity.

know what further conditions must be imposed upon the z_σ , so that the corresponding p_σ will meet the special conditions necessary for physical networks. A simple analysis of the definition (8) of z_σ , and of the well known restrictions on the p_σ , leads to the following assertion;

The physical restrictions on z_σ are exactly the same as those on p_σ .

It is obvious, for example, that conjugate complex z_σ are necessary for conjugate complex p_σ . Also, because $|z_\sigma| > 1$, z_σ dominates $-1/z_\sigma$. Then the sign of $\text{Re } p_\sigma$ is the same as that of the $\text{Re } z_\sigma$, and p_σ with negative real parts require z_σ with negative real parts, and so on.

Thus the direct choice of z_σ is restricted in exactly the same way as the choice of p_σ , except for the additional general requirement $|z_\sigma| > 1$. The last condition imposes no important restriction on the corresponding p_σ . Initially, it was adopted to make z_σ unique for any p_σ (not at a useful real frequency); but this condition does also play an essential role in the z -plane formulation of the interpolation problem.

6. NETWORK GAIN AND PHASE IN TERMS OF z

A first step in the z -plane formulation of the interpolation problem is the formulation of the network gain and phase functions, (1) and (2), in terms of z . This is most usefully examined as a transformation of functional form, rather than as a conformal mapping.

The gain and phase function (1) transforms as follows: The analytic relation between p and z is regular in the neighborhood of the singularities p_σ of the network function. Therefore, there will be similar singularities of the transformed function at the z -plane mappings of p_σ , which are z_σ and $-1/z_\sigma$. These singularities, and also suitable behavior at infinity, are exhibited by the following expression for $\alpha + i\beta$ as a function of z .

$$\alpha + i\beta = \log K'_z \frac{\prod \left(1 - \frac{z}{z'_\sigma}\right) \left(1 + \frac{1}{z'_\sigma z}\right)}{\prod \left(1 - \frac{z}{z''_\sigma}\right) \left(1 + \frac{1}{z''_\sigma z}\right)} \quad (9)$$

\prod is used here to designate a product of factors of the type following it.†

† The expression is readily confirmed in the following very elementary manner: For every factor $\left(1 - \frac{z}{z_\sigma}\right)$, in (9), there is also a factor $\left(1 + \frac{1}{z_\sigma z}\right)$. The product of the two may be expanded as follows:

$$\left(1 - \frac{z}{z_\sigma}\right) \left(1 + \frac{1}{z_\sigma z}\right) = \frac{1}{z_\sigma} \left[\left(z_\sigma - \frac{1}{z_\sigma}\right) - \left(z - \frac{1}{z}\right) \right] \quad (10)$$

If we define a new scale factor K_z by $K'_z = K_z^2$, we may write (9) as follows:

$$\alpha + i\beta = \log \left\{ K_z \frac{\prod \left(1 - \frac{z}{z'_\sigma} \right)}{\prod \left(1 - \frac{z}{z''_\sigma} \right)} \right\} \left\{ \begin{array}{l} \text{Same Rational} \\ \text{Function in } -1/z \end{array} \right\} \quad (13)$$

Similar expressions for the separate gain and phase functions may be derived from (2):

$$2\alpha = \log \left\{ K_z^2 \frac{\prod \left(1 - \frac{z^2}{z_{\sigma'}^2} \right)}{\prod \left(1 - \frac{z^2}{z_{\sigma''}^2} \right)} \right\} \left\{ \begin{array}{l} \text{Same Rational} \\ \text{Function in } -1/z \end{array} \right\} \quad (14)$$

$$i2\beta = \log \left\{ \prod \frac{1 - \frac{z}{z'_\sigma}}{1 + \frac{z}{z'_\sigma}} \prod \frac{1 + \frac{z}{z''_\sigma}}{1 - \frac{z}{z''_\sigma}} \right\} \left\{ \begin{array}{l} \text{Same Rational} \\ \text{Function in } -1/z \end{array} \right\}$$

Equation (13) holds at all values of p and z , while (14) holds at all real frequencies. Simplifications of (14) should be noted, good for the useful interval only. When $|z| = 1$, $1/z = z^*$. Recalling also that $\log |x|^2$ is $2 \log |x|$, and similar elementary relations, one obtains from (14):

When $|z| = 1$,

$$\alpha = \log \left| K_z^2 \frac{\prod \left(1 - \frac{z^2}{z_{\sigma'}^2} \right)}{\prod \left(1 - \frac{z^2}{z_{\sigma''}^2} \right)} \right| \quad (15)$$

$$\beta = \text{Phase} \prod \frac{1 - \frac{z}{z'_\sigma}}{1 + \frac{z}{z'_\sigma}} \prod \frac{1 + \frac{z}{z''_\sigma}}{1 - \frac{z}{z''_\sigma}}$$

Comparison with (5) and (7) gives:

$$\left(1 - \frac{z}{z_\sigma} \right) \left(1 + \frac{1}{z_\sigma z} \right) = -\frac{2}{\omega_c z_\sigma} (p - p_\sigma) \quad (11)$$

Thus (9) is equivalent to (1) provided

$$K'_z = K \frac{\prod \left(-\frac{z'_\sigma \omega_c}{2} \right)}{\prod \left(-\frac{z''_\sigma \omega_c}{2} \right)} \quad (12)$$

7. THE POWER SERIES IN z

Our applications to network synthesis depend upon a correspondence which may be shown to exist between certain functions of z and certain power series in z . The functions of z may be formulated in terms of network singularities. The power series in z may be derived from the Tchebycheff polynomial series in p representing the corresponding gain and phase.

The Tchebycheff polynomial expansion of a gain and phase function may be written:

$$\alpha + i\beta = \sum C_k T_k \quad (16)$$

If $\alpha + i\beta$ corresponds to a finite network, it may be represented by the function of z in (13). At the same time, T_k may be represented by the function of z in (6). With these changes, (16) becomes:

$$\begin{aligned} \log \left\{ K_z \frac{\prod \left(1 - \frac{z}{z'_\sigma} \right)}{\prod \left(1 - \frac{z}{z''_\sigma} \right)} \right\} & \left\{ \begin{array}{l} \text{Same Rational} \\ \text{Function in } -1/z \end{array} \right\} \\ & = \sum C_k \frac{1}{2} \left[z^k + \left(\frac{-1}{z} \right)^k \right] \end{aligned} \quad (17)$$

The logarithm of the product of the two rational functions, in z and $-1/z$ respectively, may be written as the sum of two logarithms. The series in sums of z^k and $(-1/z)^k$ may be written as the sum of two series. Then

$$\begin{aligned} \log \left\{ K_z \frac{\prod \left(1 - \frac{z}{z'_\sigma} \right)}{\prod \left(1 - \frac{z}{z''_\sigma} \right)} \right\} & + \log \left\{ \begin{array}{l} \text{Same Rational} \\ \text{Function in } -1/z \end{array} \right\} \\ & = \sum \frac{C_k}{2} z^k + \sum \frac{C_k}{2} \left(\frac{-1}{z} \right)^k \end{aligned} \quad (18)$$

The above expression equates the sum of two similar functions, in z and $-1/z$ respectively, to the sum of two power series, also respectively in z and $-1/z$. The theorem on which the synthesis methods are based asserts that the functions and power series in z and $-1/z$ may be equated separately, throughout the useful interval. That is:

When $|z| = 1$,

$$\log \left\{ K_z \frac{\prod \left(1 - \frac{z}{z'_\sigma} \right)}{\prod \left(1 - \frac{z}{z_\sigma} \right)} \right\} = \frac{1}{2} \sum C_k z^k \quad (19)$$

$$\log \left\{ \begin{array}{l} \text{Same Rational} \\ \text{Function in } -1/z \end{array} \right\} = \frac{1}{2} \sum C_k \left(\frac{-1}{z} \right)^k$$

The relation (18) does not, by itself, *require* (19) to be true. (19) follows from (18) if and only if the function of z has a power series expansion involving only positive powers of z , and the function in $-1/z$ has a power series expansion in $-1/z$, with the same coefficients. This added condition, however, is readily established for the useful interval.†

Combining (19) and (16) yields a most useful relationship connecting the z -plane mappings z_σ , of the network singularities p_σ , and the coefficients C_k , of the Tchebycheff polynomial expansion of $\alpha + i\beta$:

$$\alpha + i\beta = \sum C_k T_k$$

$$\sum \frac{1}{2} C_k z^k = \log K_z \frac{\prod \left(1 - \frac{z}{z'_\sigma} \right)}{\prod \left(1 - \frac{z}{z_\sigma} \right)} \quad (20)$$

In more qualitative terms:

The transformation from variable p to variable z converts an expansion in Tchebycheff polynomials in p into an expansion in a power series in z .

Thus, by working with the z_σ , in place of the p_σ , one may use a power series sort of analysis in calculating, or in choosing, the coefficients C_k in the Tchebycheff polynomial series.

The relations (20) refer to the combined gain and phase function. Exactly similar relations can readily be obtained, however, for gain and

† As defined in (8), $|z_\sigma| > 1$. In the useful interval, $|z| = 1$. Hence $|z/z_\sigma| < 1$. It follows that $\log (1 - z/z_\sigma)$ has a power (MacClauren) series expansion in positive powers of z , convergent on and within the circle $|z| = 1$. Finally the first logarithm in (19) may be expressed as a sum of logarithms of this simple type, each of which may be expanded separately. Substituting $-1/z$ for z maps the unit circle onto itself. It follows that the second logarithm in (19) has an expansion in positive powers of $-1/z$, in the useful interval, provided the first logarithm has an expansion in positive powers of z ; and the coefficients in the two series will be the same.

phase separately. These may be derived from (14), and take the form:

$$\left. \begin{aligned} \alpha &= \sum C_k T_k \\ \sum C_k z^k &= \log K_z^2 \frac{\prod \left(1 - \frac{z^2}{z_{\sigma}^{'2}}\right)}{\prod \left(1 - \frac{z^2}{z_{\sigma}^{'\prime 2}}\right)} \end{aligned} \right\} k, \text{ even}$$

$$\left. \begin{aligned} i\beta &= \sum C_k T_k \\ \sum C_k z^k &= \log \prod \frac{1 - \frac{z}{z_{\sigma}^{'}}}{1 + \frac{z}{z_{\sigma}^{'}}} \prod \frac{1 + \frac{z}{z_{\sigma}^{'\prime}}}{1 - \frac{z}{z_{\sigma}^{'\prime}}} \end{aligned} \right\} k, \text{ odd} \quad (21)$$

(The absence of factors $\frac{1}{2}$ in $\sum C_k z^k$, as compared with (20), reflects the factors 2 associated with α and β in (14).)

8. REPRESENTATION OF ASSIGNED GAIN AND PHASE

In synthesis problems, the network gain or phase, α or β , is to approximate an assigned (ideal) gain or phase, say $\bar{\alpha}$ or $\bar{\beta}$. To make effective use of the z -plane analysis, in network synthesis, we need to describe $\bar{\alpha}$ and $\bar{\beta}$ by relations analogous to (20) and (21), which express α and β in z -plane terms. These relations, while similar to (20) and (21), must take a more general form (since $\bar{\alpha}$ or $\bar{\beta}$ need only be approximately the gain or phase of a finite network). For our present purposes, the appropriate relations are those noted below.

Let $\bar{\alpha} + i\bar{\beta}$ be any function of p which has the following properties: It must be analytic throughout the useful interval. Further, there are to be no singularities within a (p -plane) distance ϵ of the useful interval, where ϵ is *finite* (but may be small). Finally, at real frequencies, $\bar{\alpha}$ and $i\bar{\beta}$ are to equal respectively the even and odd parts of $\bar{\alpha} + i\bar{\beta}$.

Under the conditions stated, $\bar{\alpha} + i\bar{\beta}$ may always be expanded in terms of our Tchebycheff polynomials T_k . Let $\sum \bar{C}_k T_k$ be the expansion. To obtain a parallel to (20), we may form (arbitrarily) a power series $\sum \frac{1}{2} \bar{C}_k z^k$. Then we may *define* a function $\bar{R}(z)$ by identifying $\log \bar{R}(z)$ with the power series. All this adds up to the following, comparable to (20):

$$\begin{aligned} \bar{\alpha} + i\bar{\beta} &= \sum \bar{C}_k T_k \\ \sum \frac{1}{2} \bar{C}_k z^k &= \log \bar{R}(z) \end{aligned} \quad (22)$$

The functions of z have the following properties: Because of the mild restrictions, which we have imposed on the singularities of $\bar{\alpha} + i\bar{\beta}$, the series $\sum \bar{C}_k z^k$ defines a function which is analytic within, and on the circle $|z| = 1$. Then $\bar{R}(z)$, also, is analytic within, and on the circle. Further, $\bar{R}(z)$ has no zeros anywhere in the same region. ($\bar{R}(z)$, however, may be more general than the rational fraction in (20).) Finally, because of the even and odd symmetries, required of $\bar{\alpha}$ and $i\bar{\beta}$, (22) may be broken into the following parallels of the equations (21):

$$\left. \begin{aligned} \bar{\alpha} &= \sum \bar{C}_k T_k \\ \sum \bar{C}_k z^k &= \log [\bar{R}(z)\bar{R}(-z)] \end{aligned} \right\} k, \text{ even}$$

$$\left. \begin{aligned} i\bar{\beta} &= \sum \bar{C}_k T_k \\ \sum \bar{C}_k z^k &= \log \left[\frac{\bar{R}(z)}{\bar{R}(-z)} \right] \end{aligned} \right\} k, \text{ odd} \quad (23)$$

In some applications, it is possible to express $\bar{R}(z)$ in closed form. In all applications, it is possible to expand $\bar{R}(z)$ as a power series, convergent in the region $|z| \leq 1$. The same is true of $1/\bar{R}(z)$, since there are no zeros in the region. Coefficients of either series ($\bar{R}(z)$ or $1/\bar{R}(z)$) may readily be calculated by means which we shall examine a little later. For the present we shall say merely that $\bar{R}(z)$ is a *known* function, corresponding to an assigned $\bar{\alpha} + i\bar{\beta}$.

9. A DESIGN CRITERION

When the gain and phase function, $\alpha + i\beta$, is to approximate $\bar{\alpha} + i\bar{\beta}$, the error in the approximation is $(\alpha - \bar{\alpha}) + i(\beta - \bar{\beta})$. The error may be expressed in terms of z by taking the difference of corresponding equations in (20), (22). The difference of the logarithms may be expressed as a single logarithm of a ratio. Alternatively, and also more conveniently for our later purposes, it may be expressed as the negative of the logarithm of the reciprocal ratio. When this is done,

$$(\alpha - \bar{\alpha}) + i(\beta - \bar{\beta}) = \sum (C_k - \bar{C}_k) T_k$$

$$\sum \frac{1}{2}(C_k - \bar{C}_k) z^k = -\log \left\{ \frac{1}{K_z} \frac{\prod \left(1 - \frac{z}{z_{\sigma}'} \right)}{\prod \left(1 - \frac{z}{z_{\sigma}} \right)} \cdot \bar{R}(z) \right\} \quad (24)$$

Consider the following arbitrary requirement, as a design criterion: The series $\sum C_k T_k$ is to match exactly the series $\sum \bar{C}_k T_k$, through

terms of order m . If both series have converged to small remainders when $k = m$, this criterion will surely make $\alpha + i\beta$ a good approximation to $\bar{\alpha} + i\bar{\beta}$.† In terms of the coefficients, the criterion requires:

$$C_k = \bar{C}_k, \quad k \leq m \quad (25)$$

If (25) is applied to the second equation of (24), the power series is zero through terms of order m . In other words, the logarithm, equated to the series, will approximate zero in the power series, or "maximally flat" manner, to order m . The logarithm is zero when the expression in brackets is unity. Further, the logarithm will approximate zero in the maximally flat manner when, and only when the bracket approximates unity in the maximally flat manner. Thus a condition which is equivalent to (25) is the following:

$$\frac{1}{K_z} \frac{\prod \left(1 - \frac{z}{z_{\sigma}'}\right)}{\prod \left(1 - \frac{z}{z_{\sigma}}\right)} \cdot \bar{R}(z) = 1 + \epsilon_{m+1} z^{m+1} + \epsilon_{m+2} z^{m+2} \dots \quad (26)$$

This may be represented symbolically by

$$\frac{1}{K_z} \frac{\prod \left(1 - \frac{z}{z_{\sigma}'}\right)}{\prod \left(1 - \frac{z}{z_{\sigma}}\right)} \cdot \bar{R}(z) \stackrel{m}{=} 1 \quad (27)$$

where $\stackrel{m}{=}$ is used to indicate equality through power series terms of order m .

When (27) is applied to network synthesis, the singularities z_{σ} , and scale factor K_z are the unknowns, while $\bar{R}(z)$ is known. If m is equal to the total number of z_{σ} , (27) will determine the network function completely. When m is smaller, (27) will furnish $m + 1$ relations between the network parameters (including the zero order condition), which may be combined with specifications of other sorts. Since (27) is equivalent to (25), this procedure amounts to the determination of network parameters which will yield assigned values of the coefficients, $C_k = \bar{C}_k$, $k \leq m$, in the Tchebycheff polynomial expansion of $\alpha + i\beta$.

Equation (27) applies when both gain and phase are to be approximated. For approximation to gain only, or to phase only, similar relations may be derived from (21) and (23). Only even ordered Tchebycheff

† When both residues are relatively large, the approximation *may* still be good, for the remainders may be quite similar, and the error will be their difference. In practical applications, this is a not uncommon situation.

polynomials contribute to gain. The following condition turns out to be the equivalent of $C_{2k} = \bar{C}_{2k}$, $k \leq m$:

$$\frac{1}{K_z^2} \frac{\prod \left(1 - \frac{z^2}{z_\sigma^2}\right)}{\prod \left(1 - \frac{z^2}{z_\sigma^2}\right)} \cdot \bar{R}(z) \bar{R}(-z) \stackrel{mc}{=} 1 \quad (28)$$

where $\stackrel{mc}{=}$ means approximation in accordance with a power series of even ordered terms, through order $2m$. Correspondingly, only odd ordered Tchebycheff polynomials contribute to phase. The following condition is equivalent to $C_{2k-1} = \bar{C}_{2k-1}$, $k = 1$ to m :

$$\prod \frac{1 - \frac{z}{z_\sigma}}{1 + \frac{z}{z_\sigma}} \prod \frac{1 + \frac{z}{z_\sigma}}{1 - \frac{z}{z_\sigma}} \cdot \frac{\bar{R}(z)}{\bar{R}(-z)} \stackrel{mc}{=} 1 \quad (29)$$

The remaining sections (except the last) develop in more detail the application of z -plane techniques to more specific synthesis problems, of various sorts. Most of these (but not quite all) are based directly on (27), (28), or (29). The exceptions use a modification of (28), in which the function of z on the left is retained, but with the zeros and poles $\stackrel{mc}{\cong}$ but not $=$. adjusted for a different kind of approximation to unity, \cong but not $=$.

In all cases, unity is approximated with one of the functions appearing in (27), (28), (29). It will be convenient to use $H(z)$ to represent the error in the approximation, or departure from unity. When gain only is of interest, the function in (28) is used, and $H(z)$ is defined by:

$$\frac{1}{K_z^2} \frac{\prod \left(1 - \frac{z^2}{z_\sigma^2}\right)}{\prod \left(1 - \frac{z^2}{z_\sigma^2}\right)} \cdot \bar{R}(z) \bar{R}(-z) = 1 + H(z) \quad (30)$$

In developing the specific techniques, we shall start with a very definite, rather special example, in order to illustrate the techniques with specific operations. This will be discussed in considerable detail in Sections 10 through 14. Thereafter we shall examine how these specific operations may be generalized, in a number of different respects.

10. AN INTRODUCTORY EXAMPLE

The example which has been chosen for detailed discussion is the equalization of the gain distortion produced by two resistance-capacity

type cut-offs. The equalization is to be accomplished with a network which has n natural modes, but no finite frequencies of infinite loss. (This is simply one of the arbitrary specifications which define this problem.)

The two cut-offs may be due to circuits or devices at two different points in a communication system, which may be represented schematically as in Fig. 5. Their effect can be described in terms of two assigned natural modes. Two assigned modes are assumed, instead of only one, because a single mode would make the problem too simple. Our present purposes will be served well, however, if we simplify the problem by requiring the two assigned modes to be *identical*, say at $p = \bar{p}_0$.

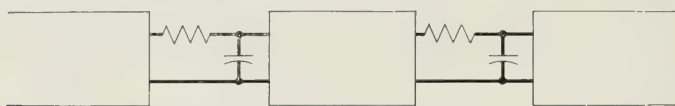


Fig. 5—A system with two resistance-capacity type cut-offs.

The two natural modes would be cancelled completely by two infinite loss points at the same location in the p plane. A network with two infinite loss points, however, is not physically possible unless it has also at least two natural modes; and the natural modes will have to introduce distortion of their own. Thus no finite network will give *perfect* equalization of unwanted natural modes. Sometimes it is desirable, in practice, to use an equalizer configuration which produces *no* finite frequencies of infinite loss, the entire equalization being accomplished by a suitable choice of its n modes. Configurations of this sort are illustrated in Fig. 6. Thus, our simple illustrative problem, though chosen to introduce principles, is also of some practical interest.

The exclusion of finite frequencies of infinite loss simplifies the repre-

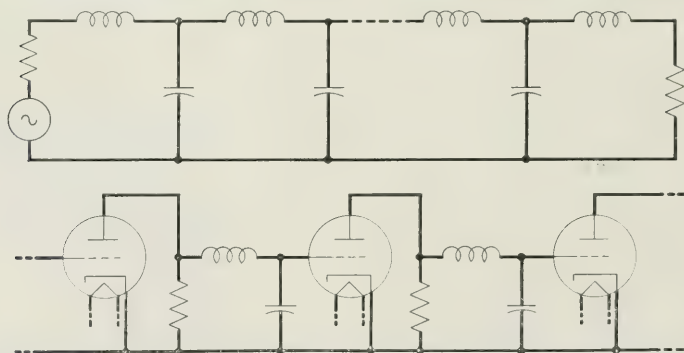


Fig. 6—Configurations which produce no finite frequencies of infinite loss.

sensation of the network gain α . In (21), the z'_σ correspond to finite frequencies of infinite loss, and are to be omitted when there are to be natural modes only. What is left is the logarithm of the reciprocal of a polynomial, which is of course the negative of the logarithm of the polynomial itself. Thus α may be described as follows, for this particular application:

$$\alpha = \sum C_{2k} T_{2k} \\ \sum C_{2k} z^{2k} = -\log K_z^2 \prod \left(1 - \frac{z^2}{z_\sigma^2}\right), \quad \sigma = 1, \dots, n \quad (31)$$

(For convenience, z_σ has been written for z''_σ , and K_z has been redefined to avoid the $1/K_z$ required if it is defined as in (21).)

The assigned gain $\bar{\alpha}$ is even more special. In this particular problem, $\bar{\alpha}$ has most of the properties of a network gain α . Specifically, it is the negative of the gain to be equalized, which in fact corresponds to a finite network. As a result, $\bar{R}(z)$ of (22) may be expressed in closed (rational) form. (Later on, we shall modify the methods appropriate for this very special situation, so that $\bar{R}(z)$ need be representable only by series.)

The specific representation of our present $\bar{\alpha}$ may be very similar to the representation of α in (31), as follows:

$$\bar{\alpha} = \sum \bar{C}_{2k} T_{2k} \\ \sum \bar{C}_{2k} z^{2k} = +\log \bar{K}_z^2 \left(1 - \frac{z^2}{\bar{z}_0^2}\right)^2 \quad (32)$$

(Both (31) and (32) apply only to the useful interval, $|z| = 1$.)

The constant \bar{z}_0 is the z -plane mapping of the assigned unwanted natural modes at $p = \bar{p}_0$, and may be calculated therefrom by (8). In (32), \bar{z}_0 determines the \bar{C}_{2k} , which in turn determine $\bar{\alpha}$. The constants z_σ , in (31), are the z -plane mappings of the arbitrary natural modes of the equalizer. They are to be adjusted to make α approximate $\bar{\alpha}$. Then the network natural modes p_σ may be calculated from them, by means of (8).

Taking the difference of corresponding equations in (31) and (32) gives the following, analogous to (24):

$$\alpha - \bar{\alpha} = \sum (C_{2k} - \bar{C}_{2k}) T_{2k} \\ \sum (C_{2k} - \bar{C}_{2k}) z^{2k} = -\log \left\{ K_z^2 \bar{K}_z^2 \left(1 - \frac{z^2}{\bar{z}_0^2}\right)^2 \prod \left(1 - \frac{z^2}{z_\sigma^2}\right) \right\} \quad (33)$$

This differs from (24) in two regards. It relates to the gain error, $(\alpha - \bar{\alpha})$,

without regard to phase. It reflects the more specific functional forms of our present α and $\bar{\alpha}$.

The formulas show that the coefficients in the Tchebycheff polynomial expansion of our present $\alpha - \bar{\alpha}$ are fixed by the logarithm of a polynomial in z^2 , of degree $n + 2$. Since the Tchebycheff polynomial series is simply one representation of the function $\alpha - \bar{\alpha}$, this means that $\alpha - \bar{\alpha}$ itself is determined by the polynomial in z^2 . Out of the $n + 2$ zeros, in terms of z^2 , n are subject to arbitrary choice, but the other two are required to be at $z^2 = \bar{z}_0^2$.

To arrive at a useful choice of the zeros, one may start with the expanded form of the polynomial, which replaces the second equation of (33) by:

$$\sum (C_{2k} - \bar{C}_{2k})z^{2k} = -\log \{ \hat{K}_0 + \hat{K}_1 z^2 + \cdots \hat{K}_{n+2} z^{2n+4} \} \quad (34)$$

All but two of the coefficients \hat{K}_k may be assigned arbitrary values, provided the remaining two are then adjusted to give the required two zeros at $z^2 = \bar{z}_0^2$. The corresponding zeros z_σ^2 may then be found by ordinary root extraction methods.

The coefficients may be chosen in such a way that the complex polynomial approximates unity, when $|z| = 1$. Then the logarithm approximates zero, the coefficients in the power series (34) are small, and since these are also the coefficients in the Tchebycheff polynomial series in (33), $\alpha - \bar{\alpha}$ is small.

11. TCHEBYCHEFF POLYNOMIAL SERIES MATCHED THROUGH n TERMS

A special choice of coefficients, which meets these requirements fairly well, is the choice determined by (28), with $m = n$. The function on the left side of (28) is here the polynomial in (34). For our present purposes, therefore, (28) becomes:

$$\{ \hat{K}_0 + \hat{K}_1 z^2 + \cdots \hat{K}_{n+2} z^{2n+4} \}^{ne} = 1 \quad (35)$$

This requires $\hat{K}_0 = 1$, and $\hat{K}_k = 0$ for $k = 1$ to n . Then \hat{K}_{n+1} and \hat{K}_{n+2} must be adjusted to give the two required zeros at $z^2 = \bar{z}_0^2$. This gives:

$$\hat{K}_0 + \cdots \hat{K}_{n+2} z^{2n+4} = 1 - (n+2) \left(\frac{z}{\bar{z}_0} \right)^{2n+2} + (n+1) \left(\frac{z}{\bar{z}_0} \right)^{2n+4} \quad (36)$$

In accordance with Section 9, this special choice of coefficients corresponds to a match of Tchebycheff polynomial series, α to $\bar{\alpha}$, through terms of order $2n$:

$$C_{2k} = \bar{C}_{2k}, \quad k \leq n \quad (37)$$

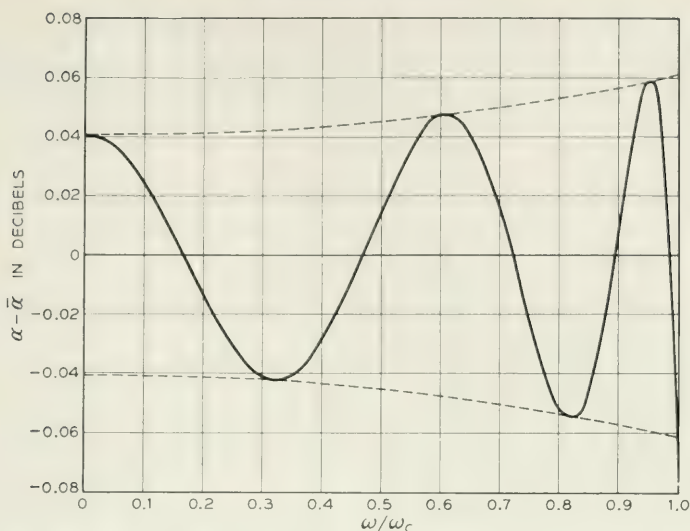


Fig. 7—Error when four natural modes equalize two natural modes at $\bar{p}_0 = -0.75\omega_c$.

The actual accuracy may be calculated from the final zeros and poles, from non-zero terms in the expansion of $\alpha - \bar{\alpha}$, or by an analysis in terms of z which will be described later.

A sample plot of $\alpha - \bar{\alpha}$ is shown in Fig. 7. This corresponds to an equalizer of four natural modes, compensating for an initial loss which rises to about 8 db at the top useful frequency, or a distortion of ± 4 db about the median loss. Residual errors are order of ± 0.06 db.

A little later we shall return to the question of accuracy, to take up methods of estimating what can be done with other numbers of arbitrary natural modes, and other values of the assigned modes. First, however, we should investigate whether the network singularities z_σ determined by (36) meet the other necessary conditions.

12. PROPERTIES OF z_σ

In the first place, $|z_\sigma|$ must be > 1 . Otherwise, the function of z in (31) will have no power series expansion over the useful interval $|z| = 1$; and (31) will not, in fact, determine the gain α over the useful interval. It turns out, however, that the condition does not give trouble in the synthesis of natural modes, when there are no arbitrary frequencies of infinite loss. This may be demonstrated by the argument outlined below.

The z_σ are zeros of the polynomial in (34), which we have given the special form (36), by applying (35). A function theoretic test for $|z_\sigma| < 1$

makes use of the contour in the complex plane for the polynomial, corresponding to the z -plane circle $|z| = 1$. (This is like a Nyquist diagram except that the contour for the variable, z , is different.) There will be $|z_\sigma| < 1$ if and only if the contour for the polynomial encloses the origin.

Now the polynomial in (34), and (35), is merely a special case of the function on the left in (28), and (30). For this special case (30) becomes

$$\sum \hat{K}_k z^{2k} = 1 + H(z) \quad (38)$$

The polynomial cannot enclose the origin without passing through some negative real value. But this requires an $|H(z)| > 1$, at some point on the contour in question, $|z| = 1$, which happens to be also our useful interval. On the other hand, $\alpha - \bar{\alpha} = 0$ when $\sum \hat{K}_k z^{2k} = 1$, and $H(z)$ is in the nature of a correction term, which is small in the useful interval when $\alpha - \bar{\alpha}$ is small.

The conclusion is: There will be no $|z_\sigma| < 1$ unless the approximation, α to $\bar{\alpha}$, is so poor that $\alpha - \bar{\alpha}$ exceeds several db in the useful interval.

Besides the requirement $|z_\sigma| > 1$, the z_σ must meet physical restrictions, which we found to be the same as those limiting the natural modes p_σ . The z_σ may be calculated as follows: The z_σ^2 are roots of the polynomial in (35), in terms of z^2 . All the roots in terms of z^2 are z_σ^2 , except the two required roots at \bar{z}_0^2 , which correspond to assigned gain $\bar{\alpha}$. Each z_σ is a square root of a z_σ^2 . There are two possible square roots, however, differing only as to sign. As far as gain α is concerned, either choice of sign is permissible; for α depends only on z_σ^2 . For a physical network, however, the choice must be such that $\text{Re } z_\sigma < 0$. This choice is possible if, and only if $\sqrt{z_\sigma^2}$ has a finite real part. A pure imaginary z_σ corresponds to a negative real z_σ^2 , and thus negative real roots in terms of z^2 are excluded by physical considerations.

Table I lists both z_σ^2 and z_σ for a number of values of n . When n is even, all roots are physical. On the other hand, when n is odd, one root is always non-physical. In a sense, an odd n is not really appropriate for the present illustrative problem, with any physical design. An odd n must necessarily bring in a real natural mode, which merely increases the sort of distortion we are trying to equalize—that is the distortion due to unwanted real modes.

The following argument substantiates the suggestion, and also illustrates manipulations of a sort which are frequently useful in more general applications: The highest order coefficient in (34), \hat{K}_{n+2} , may be set aside for adjustments to satisfy physical requirements. The rest of

TABLE I—Z-Plane Natural Modes for Equalization of Two Identical Unwanted Modes

n	z_σ^2/\bar{z}_0^2	$\sqrt{z_\sigma^2/\bar{z}_0^2}$	$z_\sigma/ \bar{z}_0 $
1	-.5000	$0 \pm i .7071$	Non Physical
2	-.3333 $\pm i .4714$	$\pm (.3492 \pm i .6747)$	$-.3492 \pm i .6747$
3	-.6059 -.0720 $\pm i .6384$	$0 \pm i .7784$ $\pm (.5340 \pm i .5977)$	Non Physical $-.5340 \pm i .5977$
4	+.1378 $\pm i .6782$ -.5378 $\pm i .3582$	$\pm (.6441 \pm i .5264)$ $\pm (.2328 \pm i .7695)$	$-.6441 \pm i .5264$ $-.2328 \pm i .7695$
5	-.6703 +.2942 $\pm i .6684$ -.3757 $\pm i .5701$	$0 \pm i .8187$ $\pm (.7157 \pm i .4670)$ $\pm (.3918 \pm i .7275)$	Non Physical $-.7157 \pm i .4670$ $-.3918 \pm i .7275$

the coefficients may then be chosen to eliminate terms from the series $\sum (C_{2k} - \bar{C}_{2k})T_{2k}$, representing $\alpha - \bar{\alpha}$, subject to the condition that two zeros must be $z^2 = \bar{z}_0^2$. This replaces $\overset{ne}{=}$ by $\overset{(n-1)e}{=}$, in (35), and changes (36) to:

$$\begin{aligned} \sum \hat{K}_k z^{2k} &= 1 - (n+1) \left(\frac{z}{\bar{z}_0}\right)^{2n} + n \left(\frac{z}{\bar{z}_0}\right)^{2n+2} \\ &+ \hat{K}_{n+2} z^{2n} (\bar{z}_0^2 - z^2)^2 \end{aligned} \quad (39)$$

If n is odd, all the roots z_σ can be physical only if \hat{K}_{n+2} is negative. On the other hand, any finite negative \hat{K}_{n+2} leads to a larger error, $\alpha - \bar{\alpha}$, than $\hat{K}_{n+2} = 0$. Reducing \hat{K}_{n+2} to zero is the same as reducing the degree of the polynomial by one, which amounts to reducing n by 1, from an odd to the next smaller even integer. In other words, a *physical* design with an odd number of natural modes is less effective, for the present application, than a simpler network, with the next smaller even number of modes.

Note that the z_σ in Table I are proportional to \bar{z}_0 . This means that root extraction methods need be used only once for each value of n , after which the roots may be quickly adjusted for any value of \bar{z}_0 , corresponding to any assigned value of the two identical modes, \bar{p}_0 .

13. ACCURACY

The accuracy of a completed design can be checked by calculating α from the natural modes p_σ , and comparing α with $\bar{\alpha}$. It is important, however, to have at least some information about accuracy in advance

of the detailed calculation of the p_σ . Otherwise, it may be necessary to carry out several designs, in all detail, in order to obtain one satisfactory design.

The needed information about accuracy can in fact be obtained from the error function $H(z)$, which we formulated for general gain applications in (30), and for the present application in (38). The analysis which yields (15) may be used to obtain a very similar expression for $\alpha - \bar{\alpha}$, in which $\bar{R}(z)\bar{R}(-z)$ appears in combination with the rational function of z from (15). It may be expressed in terms of the error function $H(z)$ of (30), as follows:

$$\alpha - \bar{\alpha} = -\log |1 + H(z)| \quad (40)$$

When $H(z)$ is zero, $\alpha - \bar{\alpha}$ is zero. When $H(z)$ is small, $\alpha - \bar{\alpha}$ depends on phase $H(z)$ as much as on $|H(z)|$. When $H(z)$ is a positive real, $\alpha - \bar{\alpha}$ is negative. When $H(z)$ is imaginary, $\alpha - \bar{\alpha}$ is very small. When $H(z)$ is a negative real, $\alpha - \bar{\alpha}$ is positive. When $H(z)$ is complex, $|\alpha - \bar{\alpha}|$ is always smaller than with a real $H(z)$ of the same magnitude. The last statement may be expressed as follows:

$$-\log \{1 + |H(z)|\} \leq \alpha - \bar{\alpha} \leq -\log \{1 - |H(z)|\} \quad (41)$$

The left hand relation is an equality when phase $H(z)$ is an even number of π radians; the right hand side, when it is an odd number of π radians.

In the useful interval, where $z = e^{i\phi}$, the $H(z)$ corresponding to (36) is as follows:

$$\begin{aligned} H(z) &= -(n+2) \left(\frac{z}{\bar{z}_0}\right)^{2n+2} + (n+1) \left(\frac{z}{\bar{z}_0}\right)^{2n+4} \\ |H(z)| &= \frac{n+2}{\bar{z}_0^{2n+2}} \left| 1 - \frac{n+1}{(n+2)\bar{z}_0^2} e^{i\phi} \right| \\ \text{phase } H(z) &= \pi + (2n+2)\phi + \text{phase} \left\{ 1 - \frac{n+1}{(n+2)\bar{z}_0^2} e^{2i\phi} \right\} \end{aligned} \quad (42)$$

As ω varies from 0 to ω_c , ϕ varies by $\frac{\pi}{2}$ radians. The corresponding phase of $H(z)$ varies by $(n+1)\pi$ radians, which means that $H(z)$ is successively positive real, imaginary, negative real, imaginary, through $n+1$ half cycles. This accounts for the oscillatory nature of the $\alpha - \bar{\alpha}$ curve, illustrated in Fig. 7.

The amplitudes of the oscillations are fixed by $|H(z)|$, which varies relatively slowly. Specifically, the two logarithms in (41) determine

envelopes, between which the actual error curve oscillates. These are the dashed lines in Fig. 7.

The maximum error, in the useful interval, is determined by the maximum value of the envelopes, i.e.,

$$(\alpha - \bar{\alpha})_{\max.} \cong \frac{n+2}{\bar{z}_0^{2n+2}} \left[1 + \frac{n+1}{n+2} \frac{1}{\bar{z}_0^2} \right] \quad (43)$$

This function is plotted in Fig. 8, for various values of n . The abscissae "distortion before equalization" represent distortion relative to the median loss in the useful interval, or one half the total variation in the interval. (This is a function of the top useful frequency ω_c , relative to the assigned natural mode \bar{p}_0 ; and (7) makes ω_c/\bar{p}_0 a simple function of \bar{z}_0 .) The figure is convenient for estimating the values of n needed for specific applications.

The various ripples in $\alpha - \bar{\alpha}$ do not all have the same amplitude, (43). For some values of n and \bar{z}_0 , the amplitudes are almost uniform; for others they are quite variable. A measure of the variability in ripple

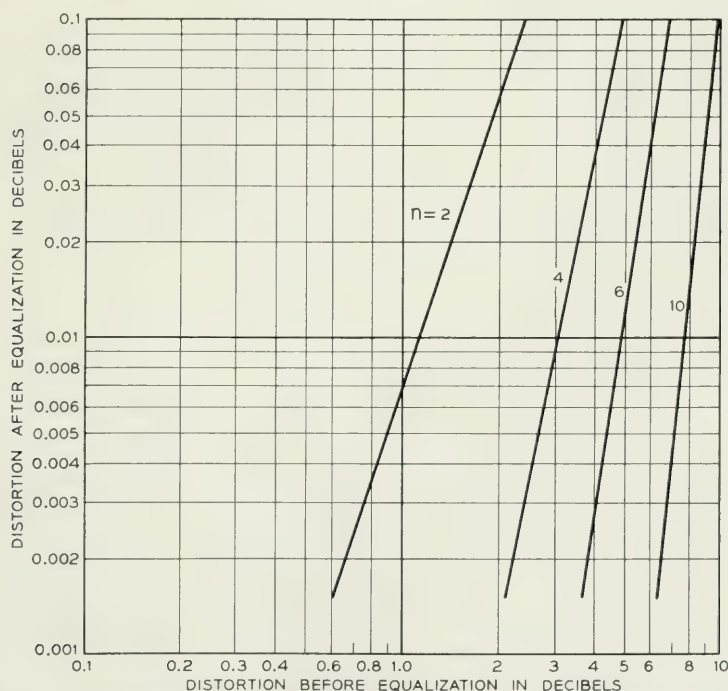


Fig. 8—Distortion before and after equalization— n natural modes equalizing 2 identical natural modes.

amplitude, across the useful interval, is:

$$\frac{|H(z)|_{\max.}}{|H(z)|_{\min.}} = \frac{(n+2)\bar{z}_0^2 + (n+1)}{(n+2)\bar{z}_0^2 - (n+1)} \quad (44)$$

14. APPROXIMATION IN THE TCHEBYCHEFF SENSE

The above analysis suggests a way of improving the design determined by (37) (or the equivalent, (36)). An *optimum* $\alpha - \bar{\alpha}$ is commonly one which has the following properties, in the useful interval:

*A maximum number of "ripples,"
all maxima of $|\alpha - \bar{\alpha}|$ equal.*

(This usually minimizes the largest departure in the useful interval, thereby yielding an "approximation in the Tchebycheff Sense.") Since the variation in phase $H(z)$ determines the number of ripples, while $|H(z)|$ determines the amplitudes of the ripples, the above conditions will be met if $H(z)$ has the following properties, in the useful interval:

*Phase $H(z)$ as variable as possible,
 $|H(z)|$ constant.* (45)

These conditions may be regarded as alternative design criteria, replacing $C_{2k} = \bar{C}_{2k}$. They can in fact be applied to our special example, and also to certain other special problems which will be noted later. For more general applications, a suitable $H(z)$ can be defined, but no reasonably simple procedure has yet been found for calculating the required constants. (The difficulties will be particularized in a later section.)

For the present example, (38) may be used to replace (34), and hence also the second equation of (33), by:

$$\sum (C_{2k} - \bar{C}_{2k})z^{2k} = -\log [1 + H(z)] \quad (46)$$

(33) requires $H(z)$ to be a polynomial in z^2 , of degree $n+2$, with two zeros of $[1 + H(z)]$ at $z^2 = \bar{z}_0^2$. The object is to find an $H(z)$ of this sort, which also satisfies (45), at least to a good approximation.

The following $H(z)$ does in fact exhibit the required properties:

$$H(z) = Gz^{2n+2} \frac{[1 - Jz^2][1 - J^{n+2}/z^{2n+4}]}{[1 - J/z^2]} \quad (47)$$

The function is a polynomial because the factor $[1 - J^{n+2}/z^{2n+4}]$ is divisible by $(1 - J/z^2)$. The constants J and G are to be chosen to

give the required double zero of $[1 + H(z)]$ at $z^2 = \bar{z}_0^2$. One value of J , so determined, is real and of order $1/\bar{z}_0^2$. This is the appropriate solution. Then $|J^{n+2}/z^{2n+4}|$ is of order $1/\bar{z}_0^{2n+4}$, when $|z| \geq 1$. This suggests the following approximation in place of (47):

$$H(z) \cong G z^{2n+2} \frac{1 - J z^2}{1 - J/\bar{z}^2} \quad (48)$$

The approximation is at least as good as $1/\bar{z}_0^{2n+4}$, compared with unity, both in the useful interval and in the neighborhood of the singularities \bar{z}_0 , and z_σ . This means that the approximation can be used: in estimating the error $\alpha - \bar{\alpha}$ (in the useful interval), in calculating J and G , and in finding the roots z_σ .

In the useful interval, $|z| = 1$, and therefore $1/\bar{z} = z^*$. Then $(1 - J/\bar{z}^2)$ is $(1 - J z^2)^*$; and their ratio has magnitude unity. Thus $|H(z)| = |G|$, in the useful interval, to order of $1/\bar{z}_0^{2n+4}$ compared with unity†. With $|J| < 1$, phase $H(z)$ varies over the useful interval to the same extent as the phase of z^{2n+2} .‡ Fig. 9 illustrates the difference in $\alpha - \bar{\alpha}$, as determined by (42) and (48). These curves, however, are for single values of n and \bar{z}_0 ; and the improvement obtained by using (48) would be different with different values of n or \bar{z}_0 .

The values of J and G , determined from (48), and the requirement that $[1 + H(z)]$ must have two zeros at $z^2 = \bar{z}_0^2$, turn out to be:

$$J = \frac{n+1}{n+2} \frac{1}{\bar{z}_0^2} \frac{2}{1 + \frac{1}{\bar{z}_0^4} + \sqrt{\left(1 - \frac{1}{\bar{z}_0^4}\right)^2 + \frac{4}{(n+2)^2 \bar{z}_0^4}}} \quad (49)$$

$$G = -\frac{1}{\bar{z}_0^{2n+2}} \frac{1 - J/\bar{z}_0^2}{1 - J \bar{z}_0^2}$$

Note that this J is in fact smaller than $1/\bar{z}_0^2$.

15. GENERALIZATION

The several sections preceding describe a quite specific example, as an introduction to synthesis applications. The next several sections describe how the specific methods of the example may be generalized, in several respects.

First, the ideal gain, $\bar{\alpha}$, is generalized, so that it need not even have the sort of functional form associated with finite networks. Then, the

† The $|H(z)|$ determined by (42) is constant only to order $1/\bar{z}_0^2$.

‡ This is the most we can expect, when we have n singularities, which can prevent the dominance of only lower order terms, through z^{2n} .

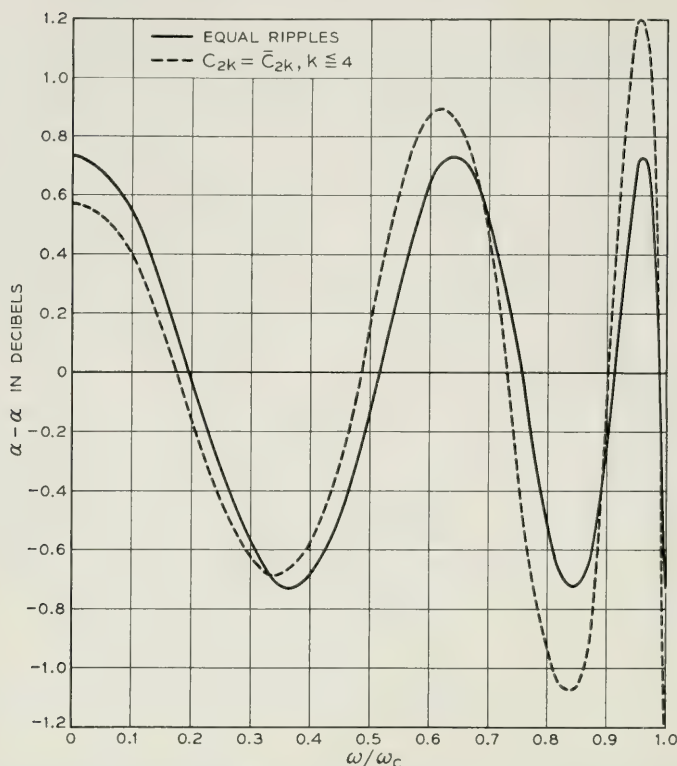


Fig. 9—Comparison of design procedures—four natural modes equalizing two natural modes at $\bar{p}_0 = -\frac{5}{12} \omega_c$.

approximating network gain is generalized, by introducing arbitrary frequencies of infinite loss, in addition to the arbitrary natural modes. The methods are also modified for approximation to an assigned phase, instead of gain, or to phase and gain simultaneously. Finally, the analysis is modified to permit useful intervals of the “high-pass” type, or (in the case of gain simulation) of the “band-pass” type.

16. APPROXIMATION TO A GENERAL ASSIGNED GAIN $\bar{\alpha}$

If we now permit the assigned gain $\bar{\alpha}$ to be general, in the sense of Section 8, we must return to the formulation:

$$\begin{aligned} \bar{\alpha} &= \sum \bar{C}_{2k} T_{2k} \\ \sum \bar{C}_{2k} z^{2k} &= \log [\bar{R}(z) \bar{R}(-z)] \end{aligned} \quad (50)$$

If we retain simulation with a network which has n natural modes, and no frequencies of infinite loss, we must retain the formulation:

$$\alpha = \sum C_{2k} T_{2k}$$

$$\sum C_{2k} z^{2k} = -\log K_z^2 \prod \left(1 - \frac{z^2}{z_\sigma^2}\right), \quad \sigma = 1, \dots, n \quad (51)$$

The corresponding formulation of the error is (in place of (33)):

$$\alpha - \bar{\alpha} = \sum (C_{2k} - \bar{C}_{2k}) T_{2k}$$

$$\sum (C_{2k} - \bar{C}_{2k}) z^{2k} = -\log \left[K_z^2 \prod \left(1 - \frac{z^2}{z_\sigma^2}\right) \cdot \bar{R}(z) \bar{R}(-z) \right] \quad (52)$$

For $C_{2k} = \bar{C}_{2k}$, $k \leq m$, the following special case of (28) is now required:

$$K_z^2 \prod \left(1 - \frac{z^2}{z_\sigma^2}\right) \cdot \bar{R}(z) \bar{R}(-z) \stackrel{me}{=} 1 \quad (53)$$

Now the reciprocal of $\bar{R}(z) \bar{R}(-z)$ has a power series expansion, in the region of interest. (Recall Section 8.)

It follows that (53) may be multiplied by this quantity, without damaging the equality of power series coefficients. In other words (53) is equivalent to:

$$K_z^2 \prod \left(1 - \frac{z^2}{z_\sigma^2}\right) \stackrel{me}{=} \frac{1}{\bar{R}(z) \bar{R}(-z)} \quad (54)$$

Let K_k be the coefficient of z^{2k} in the polynomial expansion of the left hand side; and let \bar{K}_k be the coefficient of z^{2k} in the infinite series expansion of the right hand side. Then,

$$K_z^2 \prod \left(1 - \frac{z^2}{z_\sigma^2}\right) = K_0 + K_1 z^2 + \dots + K_n z^{2n}$$

$$\frac{1}{\bar{R}(z) \bar{R}(-z)} = \sum \bar{K}_k z^{2k}, \quad k = 0, \dots, \infty \quad (55)$$

Substitution in (54) gives

$$K_0 + K_1 z^2 + \dots + K_n z^{2n} \stackrel{me}{=} \sum \bar{K}_k z^{2k} \quad (56)$$

In other words,

$$K_k = \bar{K}_k, \quad k \leq m \quad (57)$$

These relations are directly applicable to network synthesis, provided

the coefficients \bar{K}_k can be calculated. Formulae for their calculation may be derived in the following way:

If (55) is substituted in (50), the result is:†

$$\bar{\alpha} = \sum \bar{C}_{2k} T_{2k} \quad (58)$$

$$\sum \bar{C}_{2k} z^{2k} = -\log \sum \bar{K}_k z^{2k}$$

When $z = 0$, the second equation reduces to:

$$\bar{K}_0 = e^{-\bar{C}_0} \quad (59)$$

If the functions of z are differentiated, a simple rearrangement gives:

$$\sum k \bar{K}_k z^{2k-2} = [-\sum k \bar{C}_{2k} z^{2k-2}][\sum \bar{K}_k z^{2k}] \quad (60)$$

The right hand side may be expanded as a single power series, and then like powers on the two sides may be equated separately. The result is:

$$\begin{aligned} \bar{K}_1 &= -\bar{C}_2 \bar{K}_0 \\ 2\bar{K}_2 &= -\bar{C}_2 \bar{K}_1 - 2\bar{C}_4 \bar{K}_0 \\ 3\bar{K}_3 &= -\bar{C}_2 \bar{K}_2 - 2\bar{C}_4 \bar{K}_1 - 3\bar{C}_6 \bar{K}_0 \end{aligned} \quad (61)$$

Synthesis calculations may now be carried out in the following stages. The assigned gain $\bar{\alpha}$ is expanded as a Tchebycheff polynomial series, to determine coefficients \bar{C}_{2k} , say through order $k = n$. The equations (61) are then used to calculate coefficients \bar{K}_k , also through order n . Each successive coefficient is computed in terms of those previously determined. Note that the \bar{K}_k , $k \leq n$, are fixed by the same number of \bar{C}_{2k} —that is, orders $k \leq n$.

Equation (57) is now applied to identify K_k with \bar{K}_k , $k \leq m$. If all the network degrees of freedom are to be used to get $C_{2k} = \bar{C}_{2k}$, index $m = n$, and (57) determines the polynomial in (55) completely. Otherwise, $m < n$, and coefficients K_{m+1} to K_n are to be adjusted in accordance with specifications of other kinds. When all the K_k have been determined, the singularities z_σ are found by root extraction methods, applied to the right hand side of the first equation of (55).

The previous example might have been carried out in these terms, but happened to be simpler in the terms used. If (32) is regarded as a special case of (50), and if (32) is simplified (for purposes illustration) by using $\bar{K}_z = 1$, the corresponding $\bar{R}(z)\bar{R}(-z)$ becomes simply $\left(1 - \frac{z^2}{z_0^2}\right)^2$. Then $\sum \bar{K}_k z^{2k}$ is

† We may think of these equations as defining an *infinite* network, with natural modes only, which would match the assigned gain $\bar{\alpha}$ exactly.

$$\sum \bar{K}_k z^{2k} = \frac{1}{(1 - z^2/\bar{z}_0^2)^2} = \sum (k+1) \left(\frac{z^2}{\bar{z}_0^2} \right)^k \quad (62)$$

thus \bar{K}_k and K_k become

$$\begin{aligned} \bar{K}_k &= \frac{k+1}{\bar{z}_0^{2k}}, & k &= 0 \text{ to } \infty \\ K_k &= \frac{k+1}{\bar{z}_0^{2k}}, & k &= 0 \text{ to } n \end{aligned} \quad (63)$$

If these K_k are used to evaluate the polynomial on the left hand side of (53), in accordance with (55), and if the polynomial is then multiplied by the above special $\bar{R}(z)\bar{R}(-z)$, the result is exactly (36).

The error function $H(z)$, of (30) and (42), may now be defined as follows:

$$K_z^2 \prod \left(1 - \frac{z^2}{z_\sigma^2} \right) R(z)R(-z) = 1 + H(z) \quad (64)$$

The error $\alpha - \bar{\alpha}$ is again:

$$\alpha - \bar{\alpha} = -\log |1 + H(z)| \quad (65)$$

If (64) is used to express (53) in terms of $H(z)$, a "1" may be subtracted from each side of the relation, to get

$$H(z) \stackrel{me}{=} 0 \quad (66)$$

When $m = n$, this requires an $H(z)$ of the following form, in terms of the coefficients \bar{K}_k derived from $\bar{R}(z)\bar{R}(-z)$:

$$H(z) = -\frac{\bar{K}_{n+1}z^{2n+2} + \bar{K}_{n+2}z^{2n+4} + \dots}{\sum \bar{K}_k z^{2k}} \quad (67)$$

17. CHARACTERISTICS OF z_σ

As in the previous example, $|z_\sigma| > 1$ when the approximation, α to $\bar{\alpha}$, is at all reasonable. The z_σ are again zeros of $1 + H(z)$; but now $H(z)$ is defined by (64). If $|H(z)| < 1$, when $|z| = 1$, there will be the same number of zeros of $1 + H(z)$ as poles, in the region $|z| < 1$. Any poles would have to be poles of $\bar{R}(z)\bar{R}(-z)$. In Section 8, we noted that this function is regular in the region $|z| \leq 1$. Hence, there will be no poles, and there will be no $|z_\sigma| < 1$, under ordinary accuracy conditions.

As before, the z_σ can be chosen in accordance with the physical con-

ditions, provided none are pure imaginaries. Since an imaginary z_σ is a negative real z_σ^2 :

There must be no negative real z_σ^2 .

There will be no negative real z_σ^2 if the polynomial $\sum K_k z^{2k}$ in (55) is non-zero at all negative real z^2 . If the requirement is violated, initially, one or more K_k , of the highest orders, must be modified. Graphical methods are likely to be useful for this, combining plots of the original polynomial, and proposed changes. An approximation of the form $C_{2k} = \bar{C}_{2k}$, $k \leq m$, will still be realized, but with $m < n$.

The error function corresponding to $m < n$ is as follows, in place of (67):

$$H(z) = \frac{(K_{m+1} - \bar{K}_{m+1})z^{2m+2} \dots - \bar{K}_{n+1}z^{2n+2} \dots}{\sum \bar{K}_k z^{2k}} \quad (68)$$

18. ACCURACY

The accuracy of match again may be estimated by the means of (41), using $H(z)$ of (67) or (68). $H(z)$, however, may not be so easily calculated as for the previous example.

A simpler but less reliable estimate of accuracy is furnished by the error in the first unmatched coefficient in the Tchebycheff polynomial series. If $K_k = \bar{K}_k$ through $k = m$, the leading terms in various series are as follows. First, from (64) and (68),

$$K_z^2 \prod \left(1 - \frac{z^2}{z_\sigma^2}\right) R(z)R(-z) = 1 + \frac{K_{m+1} - \bar{K}_{m+1}}{K_0} z^{2m+2} \dots \quad (69)$$

using this in (52) gives:

$$\sum (C_{2k} - \bar{C}_{2k})z^{2k} = -\log \left[1 + \left(\frac{K_{m+1} - \bar{K}_{m+1}}{\bar{K}_0}\right) z^{2m+2} \dots\right] \quad (70)$$

Then, from the properties of logarithms,

$$\sum (C_{2k} - \bar{C}_{2k})z^{2k} = -\frac{K_{m+1} - \bar{K}_{m+1}}{\bar{K}_0} z^{2m+2} \dots \quad (71)$$

Consequently (also from (52)):

$$\alpha - \bar{\alpha} = \frac{\bar{K}_{m+1} - K_{m+1}}{\bar{K}_0} T_{2m+2} \dots \quad (72)$$

This is the same as the leading term of $H(z)$, except that z^{2m+2} is replaced by $-T_{2m+2}$. If $m = n$, the same equation holds with $K_{m+1} = 0$.

The coefficient $\frac{\bar{K}_{m+1} - K_{m+1}}{\bar{K}_0}$ in (72) is a sort of average of the envelopes of the ripples in $\alpha - \bar{\alpha}$. The variability of the envelopes, across the useful interval, depends upon higher order coefficients, in comparison with the leading term. Calculation of higher order coefficients is relatively complicated.

19. APPROXIMATION IN THE TCHEBYCHEFF SENSE

The criteria (45) carry over to general assigned gains, as conditions on $H(z)$ which, if realized, are usually sufficient to establish approximation in the Tchebycheff sense. For this purpose we must use the $H(z)$ of (64), rather than (67) or (68) (which correspond explicitly to $C_{2k} = \bar{C}_{2k}$, $k \leq m$). In terms of the polynomial and series representations of (55), the $H(z)$ of (64) becomes:

$$H(z) = \frac{K_0 + K_1 z^2 \cdots K_n z^{2n}}{\sum \bar{K}_k z^{2k}} - 1 \quad (73)$$

The following somewhat special problem is easily solved, in these terms, and has a direct bearing on various quite different synthesis techniques: A network is to be designed which combines the functions of an equalizer or simulator, with those of a filter, or selective network. In the useful interval, an assigned gain variation $\bar{\alpha}$ is to be approximated in the Tchebycheff sense. At higher frequencies, there is to be a rapidly increasing loss, or "sharp filter cut-off." The number of natural modes, n , is to be more than sufficient to match $\bar{\alpha}$ to the required accuracy, in the absence of a selectivity requirement, the latitude being used to produce the required sharp cut-off. In particular, n is to be large enough so that an n term match of Tchebycheff coefficients produces errors that are negligible compared with those accepted as a price of the sharp cut-off.

On the above assumption of an ample n , the infinite series $\sum \bar{K}_k z^{2k}$ in (73) may be truncated after the term of order n , and the errors due to the truncation may be neglected in calculating the design error $\alpha - \bar{\alpha}$.† Then (73) becomes

$$H(z) = \frac{K_0 + K_1 z^2 \cdots K_n z^{2n}}{\bar{K}_0 + \bar{K}_1 z^2 \cdots \bar{K}_n z^{2n}} - 1 \quad (74)$$

† The truncated series is merely the polynomial on the left side of (56) which would be obtained if the filter selectivity were ignored, and m were given the maximum value, n .

If a sharp cut-off were not required, this approximate $H(z)$ could be made exactly zero, by using $K_k = \bar{K}_k$ for all coefficients. Then the actual design error would be determined by the approximation inherent in the use of (74) in place of (73). For high selectivity, however, K_n should be much larger than \bar{K}_n , as large as possible within assigned limits on $\alpha - \bar{\alpha}$ in the useful range. (It is readily established that $K_n z^n$ will determine α at asymptotically high frequencies.) The other K_k are then to be adjusted so that $\alpha - \bar{\alpha}$ exhibits the desired "equal ripples."

The following $H(z)$ has the functional form (74), and also meets conditions (45):

$$H(z) = G z^{2n} \frac{\bar{K}_0 + \frac{\bar{K}_1}{z^2} \cdots \frac{\bar{K}_n}{z^{2n}}}{\bar{K}_0 + \bar{K}_1 z^2 \cdots \bar{K}_n z^{2n}} \quad (75)$$

Multiplying $G z^{2n}$ into the numerator gives a rational fraction which is obviously consistent with (74). The coefficients K_k of (74) which correspond to (75) are:

$$K_k = \bar{K}_k + G \bar{K}_{n-k} \quad (76)$$

In the useful interval $[\sum \bar{K}_k / z^{2k}]$ is $[\sum \bar{K}_k z^{2k}]^*$. Hence the polynomials in z and $1/z$ have identical magnitudes, in the useful interval; and, since also $|z| = 1$, $|H(z)| = |G|$ in (75). The phase variation, over the useful interval, is the same for $H(z)$ as for z^{2n} , which yields the same number of ripples in $\alpha - \bar{\alpha}$ as an ordinary Tchebycheff filter of like degree.†

The constant G is arbitrary, except that its sign must be properly chosen to avoid non-physical natural modes. Increasing G increases the filter selectivity, but also increases $\alpha - \bar{\alpha}$ in the useful interval. G and n are to be chosen together, to realize an assigned selectivity within an assigned limit on distortion.

The above analysis may be related to the following filter problem: Required to design a filter which has flat gain, in the useful interval, but which has m assigned frequencies of infinite loss, in addition to n arbitrary natural modes ($m \leq n$). The n arbitrary natural modes may be regarded as compensating for gain variations due to the assigned frequencies of infinite loss, in the useful interval, while reinforcing their effects at other frequencies. *Compensation* of effects of the infinite loss points is the same as *simulation* of the effects of natural modes at the same (assigned) frequencies. The approximation in the useful interval

† This assumes an $\bar{\alpha}$ with the general characteristics described in Section 8, which are such that the numerator and denominator of the fraction in (75) will each have a net phase shift of *zero*, across the useful interval.

is to be no better than necessary, so that there may be a maximum reinforcing of losses at other frequencies. In these terms, (58), and $\sum \bar{K}_k z^{2k}$ in (73), correspond to the assigned natural modes (at the same locations as the assigned frequencies of infinite loss). Then the ideal $\sum \bar{K}_k z^{2k}$ is itself a polynomial, of degree $m \leq n$, and (74) is exact, rather than an approximation to (73). Then (76) determines the n arbitrary modes in such a way that the net filter gain approximates zero in the Tchebycheff sense, over the useful interval.

A different procedure for obtaining the same result is described in the author's paper "Synthesis of Reactance 4-Poles".⁸ The above analysis of the filter problem is of interest in relating the more general synthesis techniques, in terms of Tchebycheff polynomial series, to previous filter theory.

Similar filters have also been obtained by Matthaei², on a potential analogy basis. He includes, however, somewhat more general filter characteristics, for which he obtains only approximately equal ripple errors. Analysis of the sort described above may be used to clarify Matthaei's analysis of the conditions under which he obtains exactly equal-ripples.

Equation (75) may be related to work of Bashkow³. The (arbitrary) amplitude of the (equal) maxima of $|\alpha - \bar{\alpha}|$, computed from $H(z)$ of (75), depends only on $|G|$. The frequencies at which the maxima occur correspond to phase $H(z) = s\pi$, which is independent of $|G|$. Thus, the locations of the maxima are invariant to the arbitrary amplitude, *within the range where (75) applies*. (75) applies only when (74) may be used in place of (73). Generally, (74) only approximates (73), and the approximation introduces small variations in the maxima of $|\alpha - \bar{\alpha}|$ (when α corresponds to (76)). If the maxima themselves are sufficiently small, the small variations will be large *percentage* variations; and the adjustments to compensate for the variations will yield significant shifts in the location of the maxima. In other words, the locations of the maxima of $|a - \bar{a}|$, required for *equal* amplitudes, are largely invariant to the magnitude of the equal amplitudes, but only to an approximation which becomes worse as the amplitudes are decreased.

Bashkow states the invariance of the frequencies of maximum $|\alpha - \bar{\alpha}|$, as a more or less empirical conclusion, based on a quite different approach to the same synthesis problem.

Equation (75) may be related also to work of Kuh.⁴ The natural modes z_σ are zeros of $1 + H(z)$. In other words, $H(z_\sigma) = -1$. Using the $H(z)$ of (75) gives the following:

$$K_0 + K_1 z_\sigma^2 + \cdots K_n z_\sigma^{2n} = -G z_\sigma^{2n} \left\{ \bar{K}_0 + \frac{\bar{K}_1}{z_\sigma^2} + \cdots \frac{\bar{K}_n}{z_\sigma^{2n}} \right\} \quad (77)$$

It must be remembered, however, that this formulation is permissible only if the approximations, inherent in (75), are justified when $z = z_\sigma$ (as well as in the useful interval). Taking the logarithm of each side gives:

$$\begin{aligned} \log \{ \bar{K}_0 + \cdots \bar{K}_n z_\sigma^{2n} \} \\ = \log (-G) + 2n \log z_\sigma + \log \left\{ \bar{K}_0 + \cdots \frac{\bar{K}_n}{z_\sigma^{2n}} \right\} \end{aligned} \quad (78)$$

Equation (58) may now be applied, to replace the logarithms by power series, provided the truncation of the series is again justified, and provided the convergence of $\sum \bar{C}_{2k} z_\sigma^{2k}$ is also proper, at both $z = z_\sigma$ and $z = 1/z_\sigma$. This gives

$$-\sum \bar{C}_{2k} z_\sigma^{2k} = -\sum \bar{C}_{2k} / z_\sigma^{2k} + 2n \log z_\sigma + \log (-G) \quad (79)$$

(Summations \sum are all with respect to k ; and there is one equation for each σ .) This is a suitable rule, for obtaining an $|\alpha - \bar{\alpha}|$ with equal maxima, whenever the approximations are in fact unimportant. It is not at all clear, however, just when the approximations become significant.

Kuh uses the potential analogy approach for the same sort of synthesis problem. He spaces the natural modes along a p -plane contour defined in fairly complicated potential analogy terms. It can be shown, however, that mapping his potentials from p plane to z plane leads to (79).

When the network is to approximate $\bar{\alpha}$ in the useful interval, but is *not* required to supply selectivity at other frequencies, the approximation (74) is usually untenable. It is generally necessary to retain the exact formulation (73).

When selectivity is not required, the phase excursion of $H(z)$, in the useful interval, can usually be increased to that of z^{2n+2} (as in (67), corresponding to $C_{2k} = \bar{C}_{2k}$, $k \leq n$). As a step toward meeting the first condition of (45), one may then write (73) as follows:

$$\begin{aligned} H(z) = -\bar{K}_{n+1} z^{2n+2} \frac{\sum \frac{\bar{K}_{n+1+k}}{\bar{K}_{n+1}} z^{2k} + \sum_1^{n+1} Q_k \frac{1}{z^{2k}}}{\sum \bar{K}_k z^{2k}} \\ Q_k = \frac{\bar{K}_{n+1-k} - K_{n+1-k}}{\bar{K}_{n+1}}, \quad k = 1, \cdots, n+1 \end{aligned} \quad (80)$$

The coefficients \bar{K}_k and $\frac{\bar{K}_{n+1+k}}{\bar{K}_{n+1}}$ are fixed by $\bar{\alpha}$. The only arbitrary

design constants are the Q_k . They are to be small enough so that they do not affect the total phase excursion $H(z)$, when $|z| = 1$. Their specific values are to make $|H(z)|$ approximately constant, when $|z| = 1$. In general the (required) series in z^2 in the numerator makes it extremely difficult to determine the required values for the Q_k . No reasonably simple general procedure has yet been found.

20. ARBITRARY RATIONAL FRACTIONS

The preceding sections were devoted to the approximation α to $\bar{\alpha}$, using n arbitrary natural modes, but no arbitrary frequencies of infinite loss. Similar techniques may be used when there are n'' arbitrary natural modes and n' arbitrary frequencies of infinite loss. As the applications become more involved, however, routine calculations must be supplemented increasingly with an element of art.

For simultaneous design of natural modes and frequencies of infinite loss, we must go back from (31) to the α formulation in (21). This we shall now write:

$$\alpha = \sum C_{2k} T_{2k}$$

$$\sum C_{2k} z^{2k} = -\log \frac{N}{D} \quad (81)$$

The functions N and D are polynomials. The coefficients will be defined as follows:

$$N = K_0'' + K_1'' z^2 \cdots K_{n''}'' z^{2n''}$$

$$D = 1 + K_1' z^2 \cdots K_{n'}' z^{2n'} \quad (82)$$

By comparison with (21), the zeros of N , in terms of z^2 , are the $z_\sigma''^2$. (Note the minus sign in 81.) The zeros of D are then the $z_\sigma'^2$. For physical networks, $n'' \geq n'$.

Equations (50), describing $\bar{\alpha}$, may be retained as they stand. Combining (50) and (81) gives, in place of (52):

$$\alpha - \bar{\alpha} = \sum (C_{2k} - \bar{C}_{2k}) T_{2k}$$

$$\sum (C_{2k} - \bar{C}_{2k}) z^{2k} = -\log \left[\frac{N}{D} \bar{R}(z) \bar{R}(-z) \right] \quad (83)$$

The function $\bar{R}(z) \bar{R}(-z)$ is exactly the same as before. The reciprocal of the function will still be $\sum \bar{K}_k z^{2k}$, with the \bar{K}_k related to \bar{C}_{2k} by (58), (59), (61). The new rational fraction N/D will appear where previously we had the polynomial $K_0 + \cdots K_n z^{2n}$.

Accordingly, the rule for $C_{2k} = \bar{C}_{2k}$, $k \leq m$, now becomes, in place of (56),

$$\frac{N}{D} = \sum \bar{K}_k z^{2k} \quad (84)$$

This condition may be used to determine the coefficients K''_k , K'_k of N and D (in combination with conditions of other sorts, when $m < n'' + n'$). When the coefficients have been calculated, the (z -plane) natural modes z''_σ may be determined from the roots of N , exactly as the z_σ of previous sections. The infinite loss points z'_σ may be calculated from the roots of D , in exactly the same way except that $\text{Re } z'_\sigma$ need not be negative. Signs of the z'_σ must be such that complex and imaginary z'_σ are in conjugate pairs. Note that there can be *conjugate* imaginary z'_σ only if D has a corresponding *double* negative real zero.

When $m = n'' + n'$, the following method may be used to calculate the K''_k and K'_k determined by (84). The relation is first multiplied by D , to get:

$$N^{(n''+n')} = D \sum \bar{K}_k z^{2k} \quad (85)$$

Then algebraic manipulation is used to evaluate the power series equivalent of the right hand side, through terms of order $n'' + n'$, using the known values of the \bar{K}_k , but general values of the K'_k of D . Each coefficient is a linear function of the unknowns, K'_k .

Now the polynomial N , in (85), has no terms of order $k > n''$. Therefore (85) requires zero coefficients in the expansion of the right hand side, from order $n'' + 1$ to order $n'' + n'$. Equating these coefficients to zero gives n' linear equations in the n' unknown K'_k . Solving for the K'_k determines polynomial D . The values calculated for the K'_k may then be used in lower order coefficients of the expansion of the right hand side of (85), which are exactly the coefficients K''_k of N .

When $n'' - n' = 0$ or 1, a continued fraction method is likely to be preferable. Various established techniques† may be used to convert the series $\sum \bar{K}_k z^{2k}$ into a continued fraction of the form:

$$\sum \bar{K}_k z^{2k} = a_0 + \frac{1}{\frac{a_1}{z^2} + \frac{1}{a_2 + \frac{1}{\frac{a_3}{z^2} + \frac{1}{a_4 \dots}}}} \quad (86)$$

† See, for example, Fry's applications of continued fractions to network design.⁹

If the continued fraction is truncated after the term of order m , and is rearranged as a rational fraction N/D , it will obey equation (84). The degrees of N and D will be such that $n'' + n' = m$, and $n'' - n' = 0$ or 1. The continued fraction may be associated with the hypothetical ladder network shown in Fig. 10, with variable impedance shunt branches proportional to z^2 . The impedance of the (truncated) ladder is N/D .

21. ACCURACY

The accuracy of match, $\alpha - \bar{\alpha}$, may again be evaluated from the final network singularities; or by (41), with $H(z)$ as in (30), before the singularities have been determined from roots of N and D . A rougher estimate may again be obtained from the error in the first unmatched term in the Tchebycheff polynomial series. As before, (equation (72)), this is equal to the leading term in $H(z)$, with z^{2m+2} replaced by $-T_{2m+2}$.

The rational fraction in (30) is the same as our present N/D . In terms of N/D , (30) becomes:

$$\frac{N}{D} \bar{R}(z) \bar{R}(-z) = 1 + H(z) \quad (87)$$

If (86), or Fig. (10), is used to determine N and D , the leading term in $H(z)$ turns out to be:

$$H(z) = \frac{(-)^{m+1} z^{2m+2}}{(a_1 a_2 \cdots a_m)^2 a_{m+1} \bar{K}_0} \cdots \quad (88)$$

The corresponding mismatch in Tchebycheff polynomial terms is:

$$C_{2m+2} - \bar{C}_{2m+2} = \frac{(-)^m}{(a_1 a_2 \cdots a_m)^2 a_{m+1} \bar{K}_0} \quad (89)$$

22. ZEROS AND POLES

When frequencies of infinite loss are to be chosen, as well as natural modes, the situation in regard to $|z_\sigma| < 1$ is somewhat less favorable.

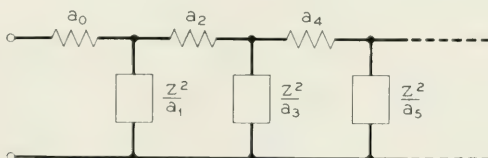


Fig. 10—A ladder network representation of the continued fraction (86).

It is still true that $1 + H(z)$ will have the same number of zeros as poles in the region $|z| < 1$, so long as $\alpha - \bar{\alpha}$ is reasonably small in the interval $|z| = 1$. In equation (87), however, the poles of $1 + H(z)$ include the zeros z'_σ of D (the arbitrary infinite loss points), as well as the poles of $\bar{R}(z)\bar{R}(-z)$. When the coefficients of D are to be chosen as in the previous section, the contour rule merely says that any z'_σ and z''_σ in the region $|z| < 1$ will occur in like numbers.

Fortunately, the frequent occurrence of $|z_\sigma| < 1$ is softened by the following curious circumstance. Almost always, any $|z'_\sigma|$ and $|z''_\sigma| < 1$ are so nearly identical that factors $(z - z'_\sigma)$ and $(z - z''_\sigma)$ may be cancelled out without any important effect on $H(z)$, or $\alpha - \bar{\alpha}$. Cancellations of this sort were encountered a number of times before an explanation was discovered. Actually the explanation is quite simple.

At any zero of $1 + H(z)$, $H(z) = -1$. On the other hand, $H(z)$ is small when $|z| = 1$. Generally, it is much *smaller* in most of the interval $|z| < 1$. For instance, when $C_{2k} = \bar{C}_{2k}$ through $k = m$, $H(z)$ is proportional to z^{2m+2} , in the neighborhood of $z = 0$. As a result, $|H(z)|$ rarely becomes as large as 1, in the region $|z| < 1$, except in the very close proximity of a pole. In other words, in the region $|z| < 1$, any zero z''_σ is usually very close to a pole z'_σ —usually so close that the corresponding factors $z - z_\sigma$ may be canceled out without significant effect on $\alpha - \bar{\alpha}$.

The occurrence of non-physical natural modes ($Re z''_\sigma = 0$) is the same as before; but adjustments to correct for these, in an efficient manner, are much more complicated. In addition, there may be non-physical infinite loss points, z'_σ . To correct for non-physical singularities, the *simplest* procedure would be to change one or both of the highest order coefficients in N and D of (82), that is $K''_{n''}$ and $K'_{n'}$. This would be inefficient, however, for it would spoil the match of $C_{n''}$ to $\bar{C}_{n''}$, or $C_{n'}$ to $\bar{C}_{n'}$. The unmodified design, defined by (84), can match terms through order $n'' + n'$, and it is desirable to change only the highest order terms in adjusting the design.

More efficient adjustments are in fact feasible. They sometimes require an increased element of art; but the art may be based on specific principles. Some particularly useful principles are described in the next section. These apply to various other corrections besides correction of non-physical zeros and poles. Examples are reduction in phase to make two-terminal realization possible, and increase in shunt capacity in two-terminal designs. In general, they offer a means of making $m < n' + n''$ in (84), and using the remaining degrees of freedom to meet other conditions.

23. MODIFICATION OF N/D

Suppose N_1/D_1 and N_2/D_2 are two rational fractions, representing special choices of N and D in (82), such that:

$$\begin{aligned}\frac{N_1}{D_1} &= \sum \bar{K}_k z^{2k} \\ \frac{N_2}{D_2} &= \sum \bar{K}_k z^{2k}\end{aligned}\quad (90)$$

Then suppose F is a function of z^2 such that

$$F^{\mu c} = 0 \quad (91)$$

The following combination represents an approximation to $\sum \bar{K}_k z^{2k}$, of the order indicated:†

$$\frac{N_1 + FN_2}{D_1 + FD_2} = \sum \bar{K}_k z^{2k} \quad (92)$$

$m = m_1$, or $m_2 + \mu$, whichever is the smaller.

The left hand side of (92) may be used as N/D in (84), to match C_{2k} to \bar{C}_{2k} through $k = m$. Adjustment of the function F may be used to satisfy other requirements, in addition to accuracy specifications.

Frequently, N_1/D_1 may be the rational fraction corresponding to truncation of the continued fraction, (86), after the term in $a_{n''+n'}$. Then N_2/D_2 is likely to be a truncation of order $m < n'' + n'$. The corresponding F is likely to be proportional to $z^{2\mu}$, or at most a simple polynomial in z^2 . When F is a constant, and $m = n'' + n' - 1$, the use of these particular rational functions in (92), to determine N/D , corresponds to matching C_{2k} to \bar{C}_{2k} through $k = n'' + n' - 1$, but leaving $C_{2(n''+n')}$ subject to adjustment. Specifically, $C_{2(n''+n')}$ depends on the choice of F , which may hinge upon such special conditions as the elimination of non-physical singularities.

Problems which call for more complicated combinations are by no means uncommon. Skill may be needed in the choice of specific combinations which will solve specific problems. Computations may be

† The relationship is easily established by noting that:

$$\frac{N_1 + FN_2}{D_1 + FD_2} - \sum \bar{K}_k z^{2k} = \frac{\frac{N_1}{D_1} - \sum \bar{K}_k z^{2k}}{1 + F \frac{D_2}{D_1}} + F \frac{\frac{N_2}{D_2} - \sum \bar{K}_k z^{2k}}{\frac{D_1}{D_2} + F} \quad (93)$$

systematized, to a considerable extent, by using the error formula (88), and other relations between the coefficients of the continued fraction, and the rational fraction truncations of various orders.†

24. APPROXIMATION TO BOTH GAIN AND PHASE

The applications described in previous sections relate to approximations to prescribed gain, $\bar{\alpha}$, without regard to the associated phase. Quite similar methods apply, however, to the simultaneous approximation of gain and phase.

The starting point is equation (20). Replacing products of factors by polynomials gives, in place of (81):

$$\begin{aligned}\alpha + i\beta &= \sum C_k T_k, & k \text{ even and odd} \\ \sum \frac{1}{2} C_k z^k &= -\log \frac{N}{D}\end{aligned}\tag{94}$$

The polynomials are now as follows, in place of (82):

$$\begin{aligned}N &= K_0'' + K_1'' z \cdots K_n'' z^{n''} \\ D &= 1 + K_1' z \cdots K_n' z^{n'}\end{aligned}\tag{95}$$

(If only natural modes are to be used, the suitable replacement for the first equation or (55) is here obtained merely by using $D = 1$, and K_z , z , z_σ , in place of their squares.)

A comparable expression is needed for the assigned gain and phase $\bar{\alpha} + i\bar{\beta}$. In place of (50), we now repeat (22), and redefine the coefficients \bar{K}_k in accordance with

$$\begin{aligned}\bar{\alpha} + i\bar{\beta} &= \sum \bar{C}_k T_k, & k \text{ even and odd} \\ \sum \frac{1}{2} \bar{C}_k z^k &= \log \bar{R}(z) = -\log \sum \bar{K}_k z^k\end{aligned}\tag{96}$$

The definition of \bar{K}_k has been changed in such a way that it is now related to $\bar{C}_k/2$ exactly as it was previously (in 58) related to \bar{C}_{2k} . Equations (59), (61) may be applied to calculating the \bar{K}_k by merely substituting therein a $\bar{C}_k/2$ for every \bar{C}_{2k} .

† For example, a simple recursion formula may be used to assemble the polynomials N and D which correspond to truncation of the continued fraction (86) at a number of different points. Specifically, $P_n = P_{n-1} + \frac{z^2}{a_n a_{n-1}} P_{n-2}$, where P is either N or D and P_n corresponds to truncation of the continued fraction after the term in a_n . The formula holds for $n \geq 2$.

Equations (84), (85), (86) may now be modified, for the new N , D and \bar{K}_k , by merely using z in place of z^2 wherever it occurs (including z^k in place of z^{2k}).† The modifications of equations (84), (85), and the truncation of (86) after a_m now lead to $C_k = \bar{C}_k$, $k \leq m$, instead of the previous $C_{2k} = \bar{C}_{2k}$. This means that m must be twice as great to match coefficients out to the same actual orders. This is to be expected since now one half of our design parameters are used to approximate phase $\bar{\beta}$, leaving only half for approximating gain $\bar{\alpha}$. Equation (89) must be changed not only in regard to the orders of C_k , \bar{C}_k , but also in regard to the factors $\frac{1}{2}$ in (94), (96). This gives

$$\frac{C_{m+1} - \bar{C}_{m+1}}{2} = \frac{(-)^m}{(a_1 a_2 \cdots a_m)^2 a_{m+1} \bar{K}_0} \quad (97)$$

The most important change is in regard to the zeros and poles z_σ . The polynomials N and D now determine z'_σ and z'_σ directly, instead of their squares. There is no opportunity to adjust the sign of $\text{Re } z''_\sigma$ by choosing the correct sign of $\sqrt{z''_\sigma}$. When non-physical singularities appear, adjustments of high order coefficients may be tried. Section 23 applies provided z^2 is replaced by z . If the specification of the problem permits added delay, linear phase may be added to $\bar{\alpha} + i\bar{\beta}$ to increase the probability of physical singularities‡. (Addition of linear phase changes only \bar{C}_1 , in $\sum \bar{C}_k T_k$. A *negative* change in \bar{C}_1 *increases* the delay.)

25. APPROXIMATION TO AN ASSIGNED PHASE $\bar{\beta}$

Sometimes it is required to approximate an assigned phase, without regard to gain. More commonly, it is required to approximate an assigned phase, using an "all-pass" network, which has a theoretically zero gain. These two problems, however, are very nearly identical, due to circumstances explained at the end of this section.

For approximation to phase only, we go back to the β equation in (21). As before, products of factors $(z - z_\sigma)$ are replaced by polynomial

† and $\overset{m}{=}$ in place of $\overset{m}{e}$.

‡ The well known relation between the gain and phase of any physical network (See for instance Bode¹⁰) may give some information regarding the reasonableness of $\bar{\beta}$. It must be remembered, however, that departures of network gain α , from the assigned gain $\bar{\alpha}$, outside the useful interval, may affect the permissible phase β , within the interval.

§ Up to the present, applications to phase problems have not been developed to the same extent as for gain. Techniques have been explored, however, to determine how such applications may in fact be carried out.

equivalents. Then, in place of (81) or (94), we have

$$\left. \begin{aligned} i\beta &= \sum C_k T_k \\ \sum C_k z^k &= -\log \frac{N}{D} \end{aligned} \right\} k \text{ odd} \quad (98)$$

Using n to represent $n'' + n'$, the total number of network singularities, we may write N and D as follows, in place of (82) or (95):

$$\begin{aligned} N &= 1 + K_1 z + K_2 z^2 + K_3 z^3 \cdots + K_n z^n \\ D &= 1 - K_1 z + K_2 z^2 - K_3 z^3 \cdots (-)^n K_n z^n \end{aligned} \quad (99)$$

Notice that N and D are related by

$$D(z) = N(-z), \quad (100)$$

which is required by the form of the β equation in (21).

To arrive at a design procedure most easily (but not the simplest design procedure), one may express the *assigned* gain $\bar{\beta}$ in the following way (comparable to (58) and (96)):

$$\begin{aligned} i\bar{\beta} &= \sum \bar{C}_k T_k, \quad k \text{ odd} \\ \sum \bar{C}_k z^k &= -\log \sum \bar{K}_k z^k \end{aligned} \quad (101)$$

Coefficients \bar{K}_k may again be calculated by a modification of (61). This time \bar{C}_{2k} is replaced by \bar{C}_k , wherever it appears in (61), and then all even ordered \bar{C}_k are made zero (since only odd terms appears in $\sum \bar{C}_k z^k$ of (101)). Note that even ordered \bar{K}_k are *not* usually zero, even though even ordered \bar{C}_k are.

The degrees of N and D , in (99), are such that we can make $C_k = \bar{C}_k$ through terms of order $k = 2n$. This requires merely:

$$\frac{N}{D} = \sum \bar{K}_k z^k \quad (102)$$

As stated, the condition applies to C_k of both even and odd orders. Since even ordered \bar{C}_k are zero, it means that at least n even ordered C_k will be zero, in addition to the match between n odd orders. (102) is sufficient to determine an N and a D without reference to (100). If the (equal) degrees n of (99) are assumed, however, the N and D determined by (102) will be found to obey (100) automatically (provided $\sum \bar{K}_k z^k$ corresponds to an *odd* series $\sum \bar{C}_k z^k$, as here assumed).[†]

A simpler method for computing the same N and D takes advantage

[†] This was discovered by Mrs. M. D. Stoughton.

of the known relation (100), connecting N and D . Let E and O be respectively the sums of even and odd terms in N . Then N is $E + O$ and (100) requires D to be $E - O$. The ratio O/E may be related to $\sum C_k z^k$ of (98) as follows:

$$\frac{O}{E} = -\tanh \frac{1}{2} \sum C_k z^k \quad (103)$$

Now let two convergent series, respectively even and odd, be such that:

$$\frac{\bar{O}}{\bar{E}} = -\tanh \frac{1}{2} \sum \bar{C}_k z^k \quad (104)$$

Let coefficients \bar{K}'_k be defined by:

$$\bar{E} + \bar{O} = \sum \bar{K}'_k z^k \quad (105)$$

Then \bar{E} and \bar{O} are respectively the sums of the even and odd terms. The complete series may now be related to the (odd) series $\sum \bar{C}_k z^k$ as follows:

$$\sum \frac{1}{2} \bar{C}_k z^k = -\log \sum \bar{K}'_k z^k \quad (106)$$

This fixes the \bar{K}'_k of $\bar{E} + \bar{O}$ in terms of the \bar{C}_k .

To make $C_k = \bar{C}_k$ through m odd orders, (102) is now replaced by

$$\frac{O}{E} \stackrel{mo}{=} \frac{\bar{O}}{\bar{E}} \quad (107)$$

The symbol $\stackrel{mo}{=}$ designates equality of power series through m odd orders. (All even terms are now zero on both sides.) The right hand side may be expressed as a continued fraction of the following form, comparable to (86):

$$\frac{\bar{O}}{\bar{E}} = \frac{1}{\frac{a_1}{z} + \frac{1}{\frac{a_2}{z} + \frac{1}{\frac{a_3}{z} + \dots}}} \quad (108)$$

Truncation after only the m^{th} term gives the O/E of (107).

The coefficients of \bar{O} and \bar{E} may be calculated by an appropriate modification of (61). (Calculate like \bar{K}_k of (101), after dividing all \bar{C}_k by 2). After O and E have been evaluated, by truncating the continued fraction (108), their sum gives polynomial N of (99).

The natural modes and frequencies of infinite loss are determined from the zeros of the polynomial N . By (21), each zero is either a (z -plane) natural mode, z''_{σ} , or the negative of an infinite loss point $-z'_{\sigma}$. If gain variations are inconsequential, there is likely to be some latitude in designating each zero as a z''_{σ} , or as a $-z'_{\sigma}$.

A zero of N with a *positive* real part would make a non-physical natural mode, and hence it *must* be a $-z'_{\sigma}$, corresponding to an infinite loss point. A zero of N with a *negative* real part *can* be a natural mode z''_{σ} , but this may not be *required*. It may be either a z''_{σ} or a $-z'_{\sigma}$, provided conjugate zeros are assigned in the same way, and provided the total number of $-z'_{\sigma}$ does not exceed the total number of z''_{σ} . The latter condition requires:

*At least half the zeros of N
must have negative real parts.*

The continued fraction (108) shows how many zeros will have negative real parts, before any zeros have been calculated. The following theorem makes this easy:

*The number of zeros of N which have negative real parts
is equal to the number of positive coefficients in the truncation of the continued fraction (108) which determined N .*

When gain is not to be disregarded, but is to be exactly zero, the synthesis technique needs few changes. The phase of an "all-pass" network is related to the natural modes z''_{σ} as follows:

$$i\beta = \sum C_k T_k, \quad k \text{ odd}$$

$$\sum \frac{C_k z^k}{2} = -\log \frac{\prod \left(1 - \frac{z}{z''_{\sigma}}\right)}{\prod \left(1 + \frac{z}{z''_{\sigma}}\right)} \quad (109)$$

This may be regarded as a special case of (20) for $\alpha = 0$ (which makes $C_k = 0$ for k even, and also happens to require $z'_{\sigma} = -z''_{\sigma}$). In functional form however, it is more like $i\beta$ of (21). It differs in only two regards. In the power series in z , each C_k is divided by two. In the rational fraction in z , all the zeros correspond to natural modes, and the poles correspond to frequencies of infinite loss; but the poles are also exactly the negatives of the zeros, as in the $i\beta$ equations of (21).

Accordingly, the phase synthesis technique which ignores gain variations may be applied to the zero gain form of the problem by cutting

every \bar{C}_k in two. All zeros of $N = E + O$ must be construed as natural modes z''_σ . Finally, the network must have as many infinite loss points as natural modes, such that $z'_\sigma = -z''_\sigma$. (Integer n is now the number of natural modes, rather than the total number of singularities.)

For physical networks, all the first n terms of the continued fraction (108) must now be positive. To meet this condition it may be necessary to add linear phase to the assigned phase (by adding a *negative* correction to \bar{C}_1). It appears that sufficient linear phase will always lead to a physical design, provided the number of modes n is increased to retain a reasonable accuracy.

26. LINEAR PHASE

When the assigned phase $\bar{\beta}$ is linear, the calculations are relatively simple.

If a delay D is to be approximated over a frequency interval extending to $\omega = \omega_c$,

$$i\bar{\beta} = -D\omega_c T_1 \quad (110)$$

If delay D is to be realized without regard to gain variations, the appropriate \bar{O}/\bar{E} is

$$\frac{\bar{O}}{\bar{E}} = \tanh \frac{D\omega_c z}{2} \quad (111)$$

A known continued fraction expansion of $\tanh X$ may be applied to (111), to obtain the coefficients of (108) without bothering with (105).† The result may be arranged as follows:

$$\frac{\bar{O}}{\bar{E}} = \frac{1}{\frac{2}{D\omega_c z} + \frac{1}{\frac{3 \cdot 2}{D\omega_c z} + \frac{1}{\frac{5 \cdot 2}{D\omega_c z} \dots}}} \quad (112)$$

Truncation of the continued fraction gives O/E , and then $O + E$. The zeros z_σ turn out to be proportional to $\frac{1}{D}$, and therefore root extraction techniques are required only for one D , for each n . The zeros are tabulated for sample n 's, in Table II.

† For the expansion of $\tanh X$, reference may be made to a text on continued fractions by Wall¹¹, page 349, equation 91.6.

TABLE II—Z-Plane Natural Modes for Linear Phase

n	$D\omega_c z_\sigma$
1	-2
2	$-3 \pm i\sqrt{3}$
3	-4.64438 $-3.67782 \pm i 3.50876$
4	$-5.79242 \pm i 1.73446$ $-4.20758 \pm i 5.31484$
5	-7.29348 $-6.70392 \pm i 3.48532$ $-4.64934 \pm i 7.14204$
6	$-8.49668 \pm i 1.73510$ $-7.47142 \pm i 5.25256$ $-5.03190 \pm i 8.98532$

The error in the first mismatched Tchebycheff coefficient is a rough measure of accuracy. It may be shown to be

$$C_{2n+1} - \bar{C}_{2n+1} = \frac{(-)^n (D\omega_c)^{2n+1}}{4^n [1 \cdot 3 \cdot 5 \cdots (2n-1)]^2 (2n+1)} \quad (113)$$

This measure of accuracy is plotted in Fig. 11, for various numbers of natural modes n . A sample detailed curve of $\beta - \bar{\beta}$ is shown in Fig. 12, with dotted lines corresponding to the estimated error (113).

If delay D is such that the error is reasonable, all the zeros may be natural modes. If these are combined with a like number of infinite loss points, such that $z'_\sigma = -z''_\sigma$, an all-pass network will be obtained, instead of one which approximates D without regard to gain. The all pass network will produce twice the delay, and twice the nonlinearity of phase. In other words, for an all pass network, both coördinates in Fig. 11 must be doubled.

27. SIMPLIFICATION OF SINGULARITY ARRAYS

In complex communication systems, a single equalizer may be required to correct for a number of effects. In a coaxial cable system, for instance, a single network in the standard repeater may be required to compensate for the following: Cable attenuation, characteristics of input and output networks, effects of interstages (significant because the feedback is limited), and distortion due to variable controls at mean settings. Tchebycheff polynomial methods may not be efficient when applied

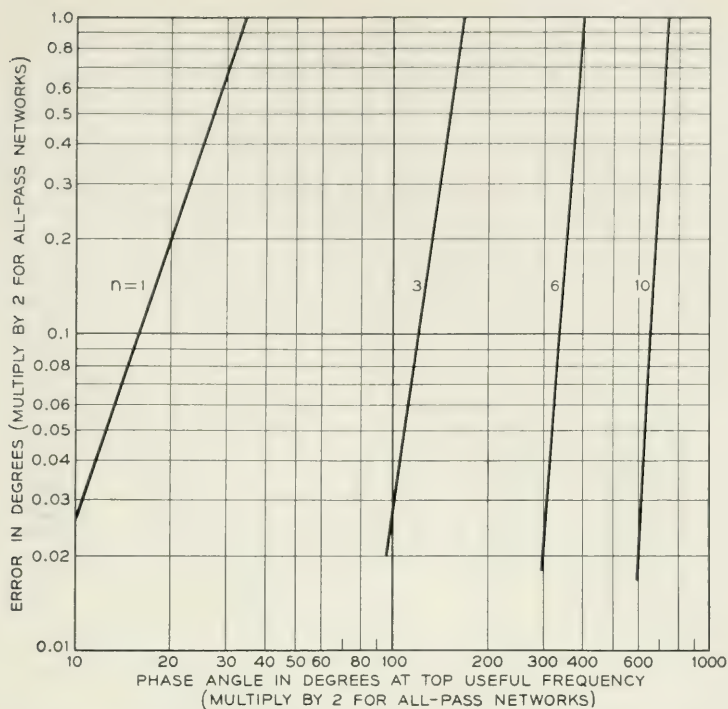


Fig. 11—Estimated error for n natural modes approximating linear phase.

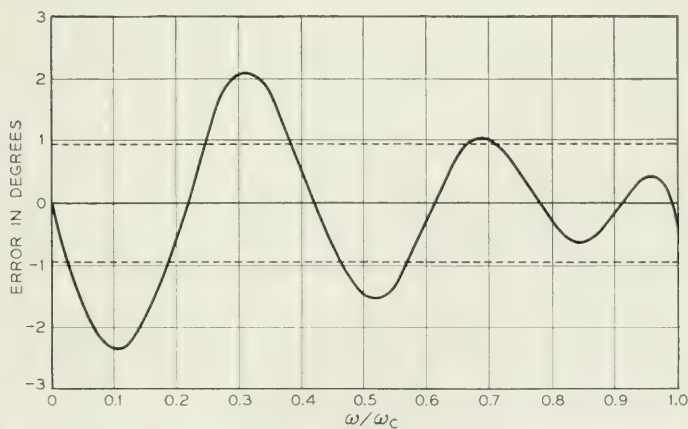


Fig. 12—Error when six natural modes approximate a linear phase, with a slope giving 402° at top useful frequency.

directly to all these effects. They may still be useful, however, when applied in the following way:

Separate arrays of singularities are determined, which match the separate effects to required accuracy, using any convenient methods. Minimum network complexity is not required at this point. Combining all the singularities in a single array gives an initial design which is sufficiently accurate, but may use many more singularities than are actually necessary. Tchebycheff polynomial methods are now used to obtain a simpler set of singularities, which approximates the initial set to sufficient accuracy. This has been designated "boiling down" the original set.

In a problem of this sort the assigned characteristic has the network form, as well as the network characteristic which is to approximate it. (The example discussed in Sections 10 to 14 is also a problem of this sort.) As a result, equations (20) and (21) apply to $\bar{\alpha}$ and $\bar{\beta}$, as well as to α and β . This makes it possible to replace $\sum \bar{K}_k z^{2k}$ and $\sum \bar{K}_k z^k$, of (55), (56), (96) etc., by a finite rational fraction \bar{N}/\bar{D} . If both $\bar{\alpha}$ and $\bar{\beta}$ are to be approximated, the following is derived from (20).

$$\sum \frac{1}{2} \bar{C}_k z^k = \log \bar{K}_z \frac{\prod \left(1 - \frac{z}{\bar{z}_\sigma'}\right)}{\prod \left(1 - \frac{z}{\bar{z}_\sigma''}\right)} = -\log \frac{\bar{N}}{\bar{D}} \quad (114)$$

The singularities \bar{z}_σ'' , \bar{z}_σ' correspond, of course, to the network singularities which are to be boiled down. If only $\bar{\alpha}$, or only $\bar{\beta}$, is to be approximated, suitable modifications are readily derived from (21).

The boiling down is accomplished by requiring

$$\frac{N}{D} \stackrel{m}{=} \frac{\bar{N}}{\bar{D}} \quad (115)$$

where N/D is of lower total degree than \bar{N}/\bar{D} . If $m = n'' + n'$, and $n'' - n' = 0$ or 1, the continued fraction method can again be used. This requires expansion of \bar{N}/\bar{D} in continued fraction form, instead of $\sum \bar{K}_k z^k$.

An example of a boiled down set of singularities is illustrated in Fig. 13.

28. GENERALIZATION OF THE USEFUL INTERVAL

All the previous analysis applies to a useful frequency interval $-\omega_c < \omega < +\omega_c$. Its important characteristics are as follows: It is a single continuous interval, with $\omega = 0$ at its center. Useful intervals with other

other characteristics may be obtained, within limits, by changing the definitions of z and z_σ , in terms of p and p_σ (equations (5) and (8)).

In all cases, the definition of Tchebycheff polynomial T_k remains the same in terms of z . The interval of orthogonality remains $|z| = 1$; and the relation between p and z is always such that the useful frequency interval is the p -plane mapping of $|z| = 1$. The relation must also be such that rational functions of p may be expressed as products of rational functions, in z and $1/z$ respectively, corresponding to (13), (14). At the same time, the physical restrictions on network singularities p_σ must translate into manageable restrictions on the z -plane singularities z_σ . It is restrictions such as these that limit the manageable useful intervals.

It is easy to apply the "low-pass" techniques to "high-pass" intervals, extending from ω_c , through ∞ , to $-\omega_c$. The appropriate trans-

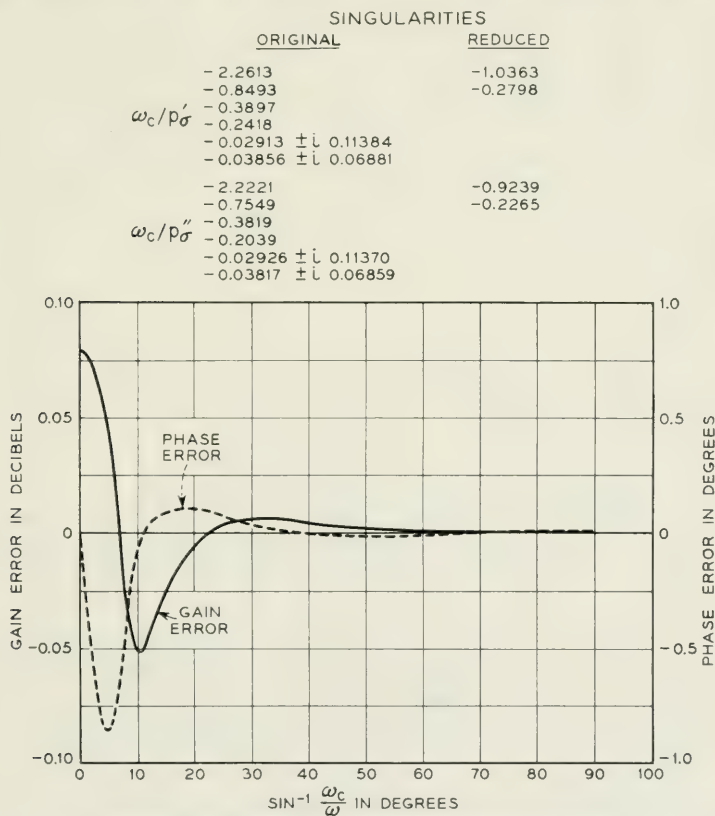


Fig. 13—An example of reduction in complexity.

formation, replacing (5), is:

$$p = \frac{2\omega_c}{z - \frac{1}{z}} \quad (116)$$

With this transformation, the whole of the previous z -plane analysis may be applied at once to high-pass useful intervals, except where linear phase is involved.

The situation is more complicated in regard to "band-pass" intervals. If the useful interval includes the frequencies between ω_{c1} and ω_{c2} , the *complete* useful interval (p -plane mapping of $|z| = 1$) must include also the "image" frequencies, between $-\omega_{c1}$ and $-\omega_{c2}$. Otherwise, conjugate complex z -plane singularities z_σ will not lead to conjugate network singularities p_σ . When there are two disjoint parts of the useful interval, the appropriate relation between p and z is relatively complicated. Up to the present, no corresponding technique has been discovered for approximating assigned phases over band-pass intervals, in Tchebycheff polynomial terms. Gain approximations can be handled, however, and for a quite simple reason. Gain functions are even functions, and behave in the p^2 plane much as gain-and-phase functions behave in the p plane. In the p^2 plane, $-\omega$ and $+\omega$ are identical, and a band-pass useful interval is a single segment of the ω^2 axis.

For gain approximations over a band-pass interval, (5) may be replaced by:

$$p^2 = \frac{a + b \left(z - \frac{1}{z} \right)^2}{1 + c \left(z - \frac{1}{z} \right)^2} \quad (117)$$

The three coefficients, a , b , c are subject to two conditions, stemming from the requirement that the interval $|z| = 1$ must map onto the interval $\omega_{c1}^2 < \omega^2 < \omega_{c2}^2$. This leaves one arbitrary degree of freedom. Its choice may be related to ordinary least squares approximations in the following way:

If $\alpha = \sum C_{2k} T_{2k}$, the first n terms approximate α in the least squares sense. In other words, the integrated square of the error is a minimum, relative to all possible choices of the first n coefficients C_{2k} , provided the integration extends over the useful frequency interval, and includes an appropriate "weight function". When (117) relates z to p , the arbitrary degree of freedom in the choice of the constants a , b , c permits selection of any one of a *family* of weight functions. Conventional

least squares analysis may be applied to determine these functions.[†]

In applying least squares analysis, it must be borne in mind that the network gain α does not approximate the assigned gain $\bar{\alpha}$ in the simple least squares sense. When $C_{2k} = \bar{C}_{2k}$ for $k \leq n$, $\alpha - \bar{\alpha}$ depends upon *two* least squares approximations. The first n terms of $\sum C_{2k} T_{2k}$ represent a least squares approximation to α , and are made identical with the first n terms of $\sum \bar{C}_{2k} T_{2k}$, which represent a least squares approximation to $\bar{\alpha}$.

When (117) relates p to z , z -plane singularities z_σ may be defined by:

$$p_\sigma^2 = \frac{a + b \left(z_\sigma - \frac{1}{z_\sigma} \right)^2}{1 + c \left(z_\sigma - \frac{1}{z_\sigma} \right)^2} \quad (118)$$

$$|z_\sigma| > 1,$$

Re z_σ to have same sign as Re p_σ

An additional singularity, z_0^2 , is also needed, corresponding to the finite poles of (117). It may be defined as follows:

$$1 + c \left(z_0 - \frac{1}{z_0} \right)^2 = 0 \quad (119)$$

$$|z_0| > 1$$

When $p_\sigma^2 - p^2$, in α of (2), is expressed in terms of z and z_σ , (117) introduces denominator factors $(1 - z^2/z_0^2)$ and $(1 - 1/z_0^2 z^2)$. As a result, α of (21) must be changed to the following, for band-pass intervals:

$$\alpha = \sum C_{2k} T_{2k}$$

$$\sum C_{2k} z^{2k} = \log K_z^2 \frac{\prod \left(1 - \frac{z^2}{z_\sigma'^2} \right)}{\prod \left(1 - \frac{z^2}{z_\sigma^2} \right)} \left(1 - \frac{z^2}{z_0^2} \right)^{n' - n'} \quad (120)$$

When definite values have been chosen for a, b, c of (117) (in order that the \bar{C}_k may be calculated), $(1 - z^2/z_0^2)$ in (120) is not subject to arbitrary adjustment. This situation can be handled by defining N/D as the rational fraction in the α equations of (21), as before, but re-

[†] For general discussions of orthogonal functions and least squares approximations, see Courant and Hilbert⁵, and also a short text by Jackson.¹²

placing (84) by

$$\frac{N}{D} \stackrel{me}{=} \left(1 - \frac{z^2}{z_0^2}\right)^{n' - n'} \sum \bar{K}_k z^{2k} \quad (121)$$

Fig. 14 illustrates an application of the technique to the simulation of a coaxial cable attenuation (which is nearly proportional to $\sqrt{\omega}$).

29. RECAPITULATION

Tchebycheff polynomial series may be applied advantageously to a very wide range of network synthesis applications. The scope of their usefulness may depend upon the skill of the designer, as with any synthesis tools, but the underlying principles are reasonably simple. The most important principles are perhaps the following:

A Tchebycheff polynomial series in frequency may be related to a power series in a new variable z . When the Tchebycheff polynomial series corresponds to a finite network gain or phase, the power series corresponds to an analytic function of z , quite similar in form to the network function of p , with singularities at z -plane mappings of the



Fig. 14—Simulation of a coaxial cable attenuation—Attenuation at top useful frequency = 46 db; Network = four constant-resistance sections.

network singularities. This makes it possible to apply power series approximation methods, in terms of z , to obtain approximations based on Tchebycheff polynomial series, in terms of frequency.

"Maximally flat" approximations in terms of z may be used to match the first m terms in the Tchebycheff polynomial series representing network gain or phase to the corresponding terms in the series representing assigned gain or phase. In this way, a Tchebycheff polynomial type of least squares approximation to the network function is made identical to the corresponding least squares approximation to the ideal function. The overall error, network function minus ideal function, is then the difference between the two least squares errors.

The z -plane analysis may also be manipulated, in a quite different way, to approach an equal ripple type of approximation (which usually represents approximation in the Tchebycheff sense). The complications are such that applications have been limited to problems of certain quite special types. On the other hand, analysis of this sort has been found useful in clarifying various other ways of seeking equal ripple approximations.

REFERENCES

1. S. Darlington, "The Potential Analogue Method of Network Synthesis," *Bell System Tech. J.*, **30**, pp. 315-365, Apr. 1951.
2. G. L. Matthaei, "A General Method for Synthesis of Filter Transfer Functions as Applied to L-C and R-C Filter Examples," *Stanford University Electronics Laboratory Technical Report No. 39*, Aug. 31, 1951 (for Office of Naval Research, NR-078-360).
3. T. R. Bashkow, "A Contribution to Network Synthesis by Potential Analogy," *Stanford University Electronics Laboratory Technical Report No. 25*, June 30, 1950 (for Office of Naval Research, NR-078-360).
4. E. S. Kuh, "A Study of the Network-Synthesis Approximation Problem for Arbitrary Loss Functions," *Stanford University Electronics Laboratory Technical Report No. 44*, Feb. 14, 1952 (for Office of Naval Research, NR-078-360).
5. R. Courant and D. Hilbert, *Methoden der Mathematischen Physik*, Vol. I, Chap. 2, Julius Springer, Berlin, 1931.
6. C. Lanczos, "Trigonometric Interpolation of Empirical and Analytic Functions," *J. of Math. and Phys.*, **17**, pp. 123-199, 1938-39.
7. H. A. Wheeler, "Potential Analog for Frequency Selectors with Oscillating Peaks," *Wheeler Monograph No. 15*, Wheeler Laboratories, Great Neck, N. Y., 1951.
8. S. Darlington, "Synthesis of Reactance 4-Poles," *J. of Math. and Phys.*, **18**, pp. 257-353, Sept., 1939.
9. T. C. Fry, "Use of Continued Fractions in Design of Electrical Networks," *American Math. Soc. Bulletin*, **35**, pp. 463-498, July-Aug., 1929.
10. H. W. Bode, *Network Analysis and Feedback Amplifier Design*, D. Van Nostrand Co., New York, 1945.
11. H. S. Wall, *Analytic Theory of Continued Fractions*, D. Van Nostrand Co., New York, 1948.
12. Dunham Jackson, *Fourier Series and Orthogonal Polynomials*, The Mathematical Association of America, Oberlin, Ohio, 1941.

A Carrier Telegraph System for Short-Haul Applications

By J. L. HYSKO, W. T. REA and L. C. ROBERTS

(Manuscript received May 14, 1952)

A compact frequency-shift carrier telegraph system is described which provides channels in the voice range and above the voice. The channel terminal unit incorporates arrangements for handling TWX supervisory signals and employs no electro-magnetic relays.

INTRODUCTION

Most short Bell System telegraph circuits, particularly those in the less-densely populated areas of the country, have customarily been operated over direct-current facilities obtained by compositing or simplexing physical telephone circuits. Many of these extend from a telegraph repeater in a central office to another arranged as a subscriber set and mounted in the knee-well of the customer's teletypewriter table. Thus, for example, circuits are extended to Teletypewriter Exchange Service (TWX) subscribers located far from the switchboard. The TWX facilities are arranged to handle supervision as well as transmission. The form of supervision is identical to that obtained when local facilities are employed and hence uniform operating procedures are obtained at TWX switchboards for all subscriber stations without regard to their geographical location.

During and immediately following World War II, the growth of the Bell System's telegraph business resulted in some shortage of dc facilities. It was foreseen that this shortage would be rapidly intensified by the use of new short-haul carrier telephone systems, such as type N1,¹ in providing telephone circuits without adding physical conductors. Moreover, many of the existing direct-current facilities would be absorbed to meet signaling needs for the rapid expansion of telephone toll dialing. It therefore became evident that carrier telegraph methods must be adopted for relatively short hauls in fringe areas.

The existing 40C1 voice-frequency carrier telegraph system^{2, 3} was

designed for application in large groups at telegraph central offices and for trunk-service operation over toll telephone circuits employing standard levels. It has proved very economical in this field. However, the very features which make for economy in large installations (such as amplitude modulation, common carrier supply and testing equipment, and standardized operating conditions) cause this equipment to be costly when it is applied a few channels at a time in outlying offices; these may not be equipped with either telegraph battery supplies or telegraph boards. Moreover, the 40C1 equipment, being a carrier-on-for-mark and carrier-off-for-space system, does not lend itself to the provision of TWX toll subscriber line supervision identical to that of local stations without the addition of rather complex and expensive supervisory applique circuits. Where TWX supervision is involved these supervisory circuits are required to generate and recognize supervisory signal patterns capable of being distinguished from transmission space signals and communication breaks, which are long space signals.

Consequently, it was decided to develop a new carrier telegraph system especially aimed at the needs of fringe areas. One of the problems to which much thought was given concerned the choice between amplitude-modulation and frequency-shift operation. A frequency-shift system provides some reduction in the effect of noise and other interference on transmission and it is also less affected by rapid level changes. Although these advantages were attractive, it was not clear that they were sufficient to justify the added complexity and cost entailed by the adoption of this type of transmission, in view of the quiet and stable circuits encountered in the Bell System plant. What finally swung the balance to a frequency-shift system was its advantage in handling TWX supervisory signals. With transmission accomplished by shifting the carrier frequency, supervisory signals could be sent by turning the carrier on and off. A cheap and simple circuit might then be used to distinguish between transmission and supervision.

From the foregoing discussion it will be evident that during the twelve years since the 40C1 system was developed the needs of the Bell System have changed. Fortunately, the designer's art has concurrently made great strides in making available new miniature apparatus and electronic techniques such as have been exploited so successfully in the 143A type electronic telegraph regenerative repeater,⁴ the V3 telephone repeater⁵ and the N-1 carrier telephone system. As a result, the channel terminal of the new 43A1 carrier telegraph system, being small, inexpensive, self-contained and all-electronic with no electro-mechanical

relays, is almost ideally suited to the needs of the smaller central offices and TWX stations.

FREQUENCY ALLOCATIONS

The 43A1 system provides two groups of channel-frequency allocations, as follows:

a—A three-channel high-frequency allocation, using frequencies between the upper edge of the voice-frequency band and the lower edge of the type-C carrier telephone band. This allocation is primarily for operation on open-wire lines but can also be operated on cable circuits where the loading provides a suitably high cut-off.

b—A voice-frequency allocation capable of providing six channels on two-wire circuits or twelve channels on four-wire circuits. The channels of this allocation are for operation over telephone speech channels on any of the standard facilities, including broad-band carrier and cable or open-wire physical circuits.

The present frequency allocations are shown in Fig. 1. The voice-frequency system is based on twelve nominal midband frequencies spaced 170 cycles apart from 595 cycles to 2635 cycles, omitting 1615 cycles. The carrier frequency is shifted ± 35 cycles about midband, and either the higher or the lower frequency may be used for marking sig-

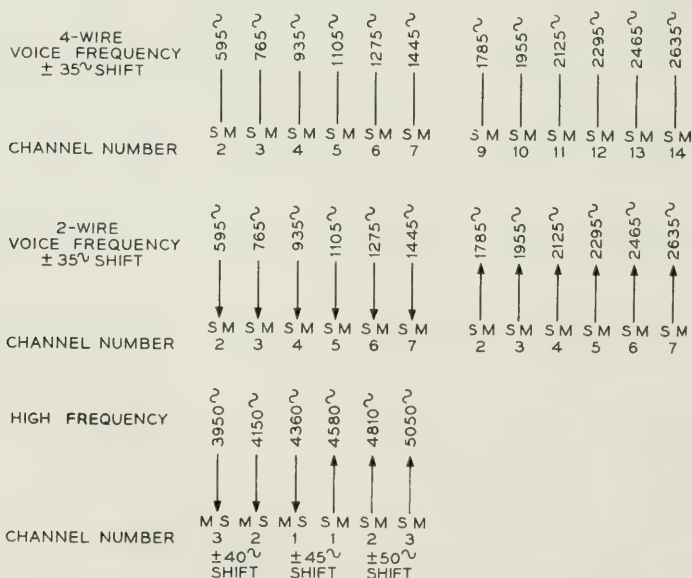


Fig. 1—Frequency allocations.

nals. The high-frequency system is based on six midband frequencies spaced from 200 to 240 cycles apart. The frequency shift ranges from ± 40 cycles in the lowest channel to ± 50 cycles in the highest. These wider spacings and shifts were adopted to ease the problem of designing inexpensive filters and oscillators for the higher frequencies.

Channel 1 of the high-frequency system employs adjacent frequency assignments for the two directions of transmission. The lower frequency path employs a downward shift for marking signals and the higher-frequency path an upward shift for marking signals. In half-duplex operation this minimizes interference from the strong signals at the transmitter output to the weak signals at the receiver input. The steady marking frequency, which is being sent against the flow of traffic, is shifted away from the band over which the message is passing.

CHANNEL TERMINAL CIRCUIT

Sending Circuit

The sending portion of the channel terminal circuit is shown in the upper part of Fig. 2. When the teletypewriter sending contacts open the loop to send a spacing signal, the sending triode is cut off. When the contacts close the loop to send a marking signal, the grid is made positive with respect to the cathode, the tube conducts and the potential at the plate is decreased. A varistor bridge modulator is connected between this plate and a source of potential having a value lying midway between the marking and spacing plate potentials. Thus, the potential applied across the modulator during marking signals is opposite in polarity to that applied during spacing signals. When the voltage across the varistor bridge is in the conducting direction, the varistors provide a low impedance path to alternating currents. Thus additional capacitance is coupled to the oscillator tank circuit and the oscillator operates at the lower of its two signal frequencies. When the voltage across the bridge is in the non-conducting direction, the varistors are biased to a high-impedance portion of their characteristic, the capacitor is effectively disconnected from the tank circuit and the oscillator operates at its higher signal frequency.

The reversing switch in the driving circuit of the modulator permits either the higher or the lower frequency to be used for marking signals.

The oscillator output power is adjusted by the SEND LEVEL potentiometer and passed through a buffer amplifier and a band pass filter to the send bus and line.

The filter is an impedance transforming structure which contains a

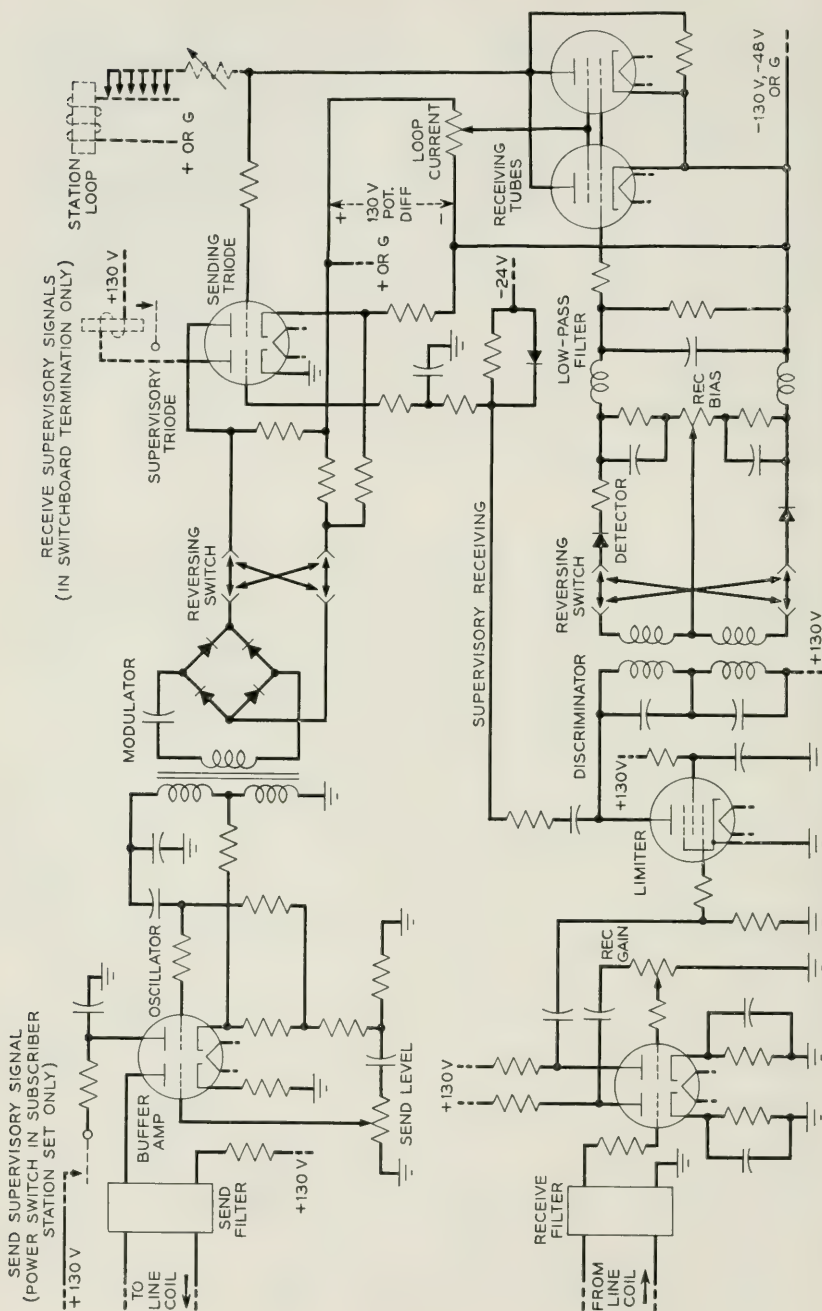


Fig. 2—Simplified diagram of channel terminal circuit, half-duplex.

downward transformation (7500:600) from the impedance of the buffer amplifier to that of the line so that no amplifier output transformer is required.

Either unity-ratio line coils or a hybrid coil may be used to connect the unbalanced sending and receiving filters to a balanced line. The hybrid coil is used with a two-wire line when the send and receive frequencies occupy adjacent bands.

Receiving Circuit

The receiving circuit, shown in the lower part of Fig. 2, is equipped with a filter which selects a narrow band of frequencies centered about the mark and space frequencies of the channel to be received. The receiving band filter has characteristics similar to those of the sending filter, except that it has a greater discrimination against unwanted frequencies and provides an upward transformation (600:140,000) from the line impedance to a value suitable for driving the grid of the first amplifier stage.

The frequency-loss characteristics of a typical receiving filter used in the voice band and of the corresponding sending filter are given in Fig. 3.

The carrier signals selected by the receiving filter are passed through a three stage amplifier limiter. Most of the limiting action is provided by the third stage; the first and second stages act as amplifiers only for weak signals but limit strong signals. An adjustment for receiving gain

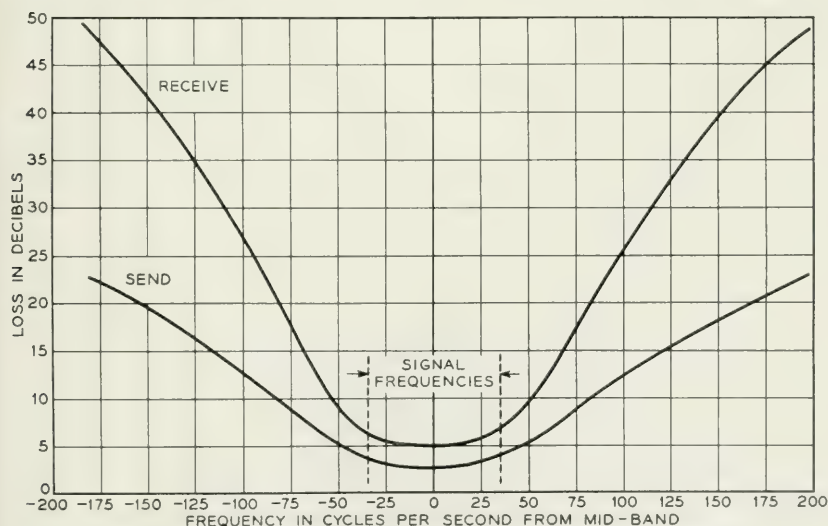


Fig. 3—Send and receive filter characteristics, VF allocation.

is provided between the first and second stages. With the REC GAIN control at maximum the third stage is driven to full output when the input to the receiving filter is greater than about -50 dbm*. Where it is not necessary to detect the presence or absence of carrier for supervision as described below, the control is generally set for maximum gain.

The limiter output is passed through a frequency discriminator consisting of two anti-resonant circuits in series, tuned so that one has a parallel resonance at the low frequency edge and the other at the high frequency edge of the channel band. The voltages appearing across the anti-resonant circuits are rectified separately by germanium varistor diodes and the resultant d-c output voltages are added algebraically, filtered and applied to the control grids of the output tubes.

Since at normal receiving levels the limiter removes all magnitude variations, the output from the discriminator detector circuit is dependent in magnitude and sign only on the signaling frequency. A negative voltage from the detector causes cut-off of the amplifier tubes and a positive voltage causes plate current to flow. A switch between the discriminator and the detector provides means for reversing the output connections of the discriminator so that a positive voltage from the detector can be obtained with either the higher or the lower signaling frequency. Fig. 4 shows the dc voltage output versus frequency characteristic obtained with a typical discriminator.

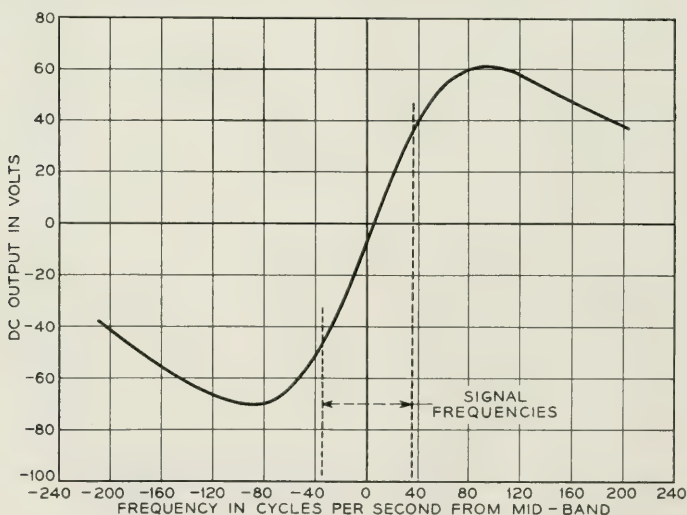


Fig. 4—Frequency characteristic of typical discriminator.

* "dbm" is an abbreviation for "decibels with respect to one milliwatt."

The low pass filter between the detector and last stage serves to remove carrier ripple and to decrease the effects of noise and other interference having demodulated frequency components greater than about 40 cps, which is slightly higher than the "dotting" frequency of 100-word per minute signals. In order to prevent a change in the tuning of the discriminator when the reversing switch is operated, a balanced low-pass filter structure without mutual inductance is employed. This presents high and nearly equal impedances to ground from the positive and negative sides of the detector circuit.

Nearly all the voltage gain of the receiver appears ahead of the detector. Since the detector output voltage applied to the grids of the beam power tetrodes is high enough to give an approximately square signal wave shape in the loop, no intermediate stage of dc amplification is needed following the detector. For unbiased signal reception, the demodulated signals should be centered on the grid characteristic of the receiving tubes; that is, the marking and spacing voltages applied to the grid circuit should be symmetrical about a potential a few volts negative with respect to the receiving tube cathodes. To obviate the need for a voltage source negative with respect to the cathodes, the signals are prebiased by unbalancing the detector so that the mean of the mark and space output voltages from the low pass filter is about -5 volts. Further adjustment of the mean signal value may be made by means of the REC BIAS potentiometer to compensate for bias of signals received from the line due to deviations in the mark and space frequencies from their theoretical values or to other causes originating at the sending terminal of the telegraph circuit as well as for bias due to discrepancies in the discriminator network or to differences between mark and space levels.

These arrangements permit great freedom in the assignment of loop battery voltages. The cathodes of the final stage may be fixed at -130 -volt, -48 -volt or ground potential and the plates operated from ground, $+48$ -volt or $+130$ -volt potential. The remainder of the circuit may be powered by $+130$ -volt battery for the plates and -24 -volt battery for the heaters of the tubes, regardless of the loop conditions.

By means of the reversing switch mentioned above, current may be caused to flow in the loop during the reception of the higher or the lower frequency. Thus not only can various frequency allocations be accommodated, but the local circuit may be operated neutral (current for mark) or inverse neutral (no current for mark).

One tube is used in the final stage for 20 ma or 30 ma loop current, and two for 60 ma loop current.

Supervisory Circuit

When the channel is used in TWX service as a toll subscriber line, the subscriber calls the operator to initiate a call by closing the power switch on his teletypewriter. This connects power to the teletypewriter motor, closes the transmission circuit to the teletypewriter and applies plate battery voltage to the transmitting oscillator in the channel terminal, resulting in the transmission of carrier current over the line. At the distant (switchboard) terminal the receipt of carrier current energizes a supervisory signal receiving circuit which is responsive to carrier-on and carrier-off conditions in the receive band. In this circuit, carrier voltage appearing at the plate of the limiter tube is rectified and applied to the grid of the supervisory triode. The operation of a relay in the plate circuit of this tube causes a line lamp at the switchboard to light.

A disconnect signal is sent by the subscriber at the end of a call by opening the teletypewriter power switch. This removes the oscillator plate voltage. At the central office, the receipt of the resulting no-carrier signal de-energizes the supervisory receiving circuit and causes the supervisory lamp in the operator's cord circuit to light steadily. To recall the operator during a call the subscriber opens and recloses his power switch. This causes the cord lamp at the switchboard to flash.

An RC circuit slows the rise of current in the supervisory receiving tube to guard against false operation of the switchboard line lamp due to noise impulses during the carrier-off, that is, the idle condition.

DC Circuits

On the dc side of the channel terminal, provision is made for optional wiring arrangements to connect to the circuits of the various telegraph test boards, service boards and TWX switchboards, as well as to local teletypewriter loops, using telegraph voltages of either 130 or 48 volts. In offices where a negative 130-volt battery is not provided, operation with a single positive 130-volt battery is possible.

The loop connections are made to an electronic circuit in the channel terminal which is similar to that employed in a recently-developed electronic loop repeater used in telegraph offices and which possesses several interesting features. Fig. 5 compares the action of this circuit, in transmitting toward the subscriber station, with that of more conventional arrangements:

(a) shows a conventional open-and-close circuit and the wave shapes which it produces at the central office end and at the far end of a capaci-

tative loop. As is well known, the asymmetrical wave shape causes positive signal bias.

(b) shows an "effective polar" circuit along with the wave shapes it delivers. This is the circuit conventionally used to drive subscriber loops. It presents a constant low impedance to the loop and might therefore be considered a "close-and-close" circuit.

(c) shows the electronic loop circuit. The driving tetrodes are operated in their high-impedance region, above the knee of the plate-current voltage curve. They deliver a highly symmetrical rectangular wave to the loop and little or no bias results. This circuit presents a nearly constant high impedance to the loop and might be considered an "open-and-open" circuit.

Although the rectangular wave is inferior to a peaked wave in that less average power is delivered for the same values of steady-state current and voltage, it provides entirely acceptable transmission for 19-gauge cable loops up to about 20 to 25 miles in length. Inasmuch as 80 volts potential is absorbed in the electron tube plate circuits, this is almost the maximum length over which 62.5 ma can be supplied when loop battery of 260 volts is used.

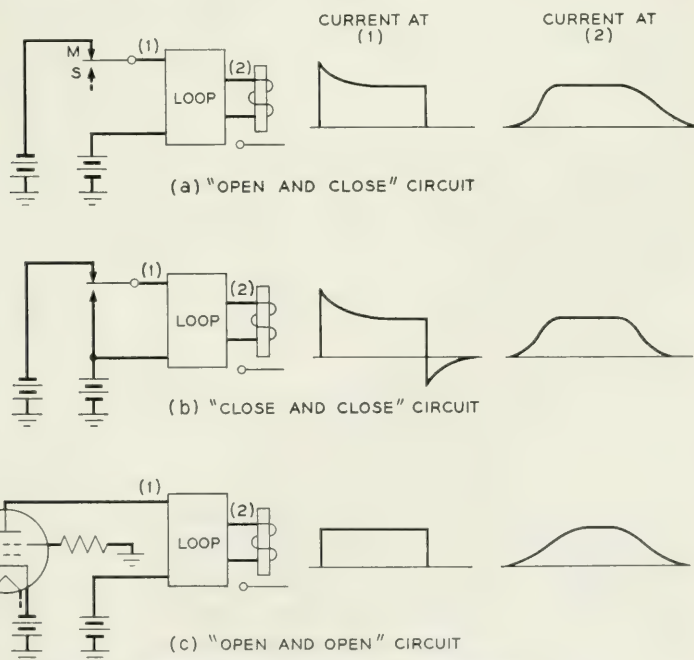


Fig. 5—Explanation of electronic loop circuit.

During open-and-close transmission by the subscriber, the high impedance termination of the loop at the tetrode plate circuits causes the current at the central office end of the loop to change very slowly—too slowly for good transmission at teletypewriter speeds. However, the voltage wave is very well shaped, and this is what is used to drive the grid of the transmitting tube. One noteworthy fact is that the bias of the signals repeated from loop to line is nearly independent of loop length; consequently no inductive wave shaping is required at the subscriber station, even in the longest operable loops.

Because of the high impedance termination, loop current is insensitive to circuit resistance. The loop padding rheostat is, therefore, adjusted to build out the loop resistance to a standard value and the amount of loop current required for proper operation of the station teletypewriter is obtained by varying the screen grid potential of the tetrode tubes.

Duplex Feature

In half-duplex operation, one dc loop at each channel terminal serves for both sending and receiving. The central office end of this loop is connected to the grid of the sending triode and to the plates of the receiving tetrodes. If a marking signal is being received from the carrier line while the teletypewriter contacts in the loop are closed, the receiving tubes conduct, current flows in the loop and the teletypewriter in the loop receives a marking signal. Under this condition the office end of the loop is positive with respect to the cathode of the sending triode; hence this tube passes a marking signal toward the carrier line. When a spacing signal is received from the carrier line the tetrodes are cut off, the loop current is reduced practically to zero, the teletypewriter receives a spacing signal and the voltage at the office end of the loop becomes more positive. Hence a marking signal continues to be transmitted to the line during the receipt of either mark or space signals from the line.

When the subscriber opens the loop to send a spacing signal to the distant terminal, the potential of the sending triode grid becomes negative with respect to its cathode, the tube cuts off and hence, as described previously, a spacing frequency is passed to the ac line.

In full-duplex operation, two loops are provided at each channel terminal to permit sending and receiving simultaneously. The grid of the sending tube is disconnected from the plates of the tetrodes and transferred to a resistive connection which terminates the full-duplex sending loop. The loop circuits operate in the same way as described for half-duplex operation except that no break action is provided.

Break Feature

When the subscriber opens the loop at the teletypewriter to break transmission coming from the distant terminal, a clean-cut space should be transmitted to the line regardless of any incoming signals. The resistor shunted between the plates and cathodes of the receiving tubes causes the central office end of the open loop to assume the same potential as the tetrode cathodes. This insures that a steady spacing potential will be applied to the send tube even though the tetrodes are cut off by an incoming space. This provides a rapid, clean break. However, if a large leakage exists across the loop conductors, the resistor will not be able to keep the sending tube in a cut-off condition and a break by the subscriber will result in the incoming transmission being reflected in an inverted condition to the distant carrier terminal. In such a case the distant sending subscriber would be broken by a "bust-up" of local copy or by operation of the keyboard break lock. This would normally be caused only by a trouble condition in cable loops.

If a break signal is received over the line from the distant end while the near end subscriber is sending, his loop current is reduced to practically zero. This operates the keyboard break lock thus breaking the subscriber. This circuit differs from the conventional loop circuit in that the receipt of a break signal does not stop the outgoing signals except via the break lock.

TELEGRAPH DISTORTION

On quiet circuits, total distortion per section averages 1 to 2 per cent at 60 words per minute and about 5 per cent at 100 words per minute. Plots of received signal distortion versus level of received carrier are shown on Fig. 6 for both signaling speeds.

EQUIPMENT FEATURES

The channel terminal employs a formed sheet-metal framework and occupies a space $10\frac{1}{2}$ inches high, $5\frac{1}{4}$ inches wide and $7\frac{3}{4}$ inches deep overall. Fig. 7 shows a 43A1 channel terminal. It is plug-terminated, and hence removable for maintenance or repair at a bench.

The basic portion of the channel terminal is common to all frequency allocations. The oscillator network and send filter, which constitute the elements determining the transmitted frequency, form a plug-terminated sub-assembly $7\frac{3}{4}$ inches high, $5\frac{3}{8}$ inches wide, and $1\frac{1}{2}$ inches deep. The receive filter and discriminator, which select the received frequency, form a plug-terminated sub-assembly of the same size.

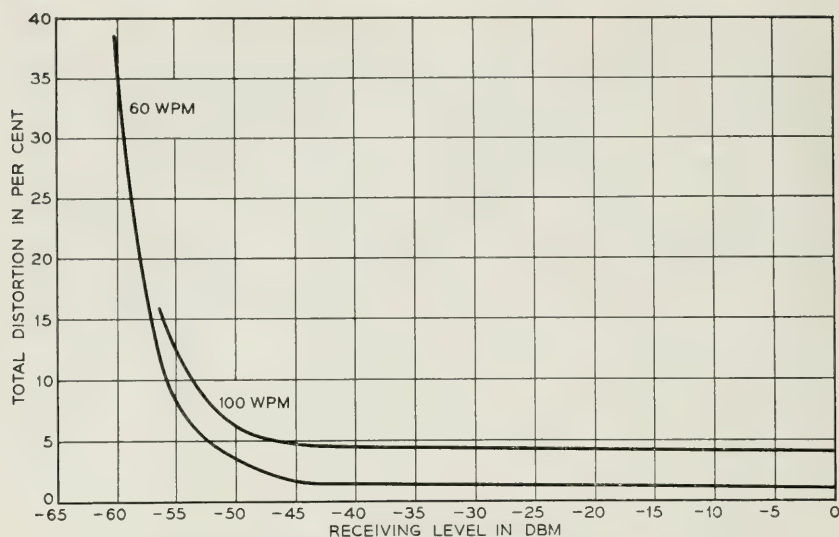


Fig. 6—Telegraph distortion vs receiving level.

A rear view of the channel terminal with the send frequency unit removed is shown in Fig. 8. With both frequency units in place, the rear of the channel terminal is almost completely enclosed. When they are removed, the wiring and apparatus terminals of the basic channel terminal are readily accessible for test and repair.

Tube sockets, potentiometers, test points, switches and the inductor of the low-pass filter are mounted on the front panel. Small resistors, capacitors, and germanium diodes are assembled on a plastic "ladder" which is mounted vertically in the space between the frequency units.

As shown in Fig. 9, three channel terminals may be mounted abreast on a welded metal frame which is fastened to any of the standard bay frameworks designed to accommodate 19-inch mounting plates. The unit mounting frame carries the multicontact receptacles into which the channel terminals are plugged. Twenty-four channel terminals may be mounted on an 11½-foot relay rack, with line coils and certain auxiliary equipment.

Where arrangements for switching between half and full-duplex operation are required, duplex switches for a number of channel terminals are mounted on a narrow plate between the channel terminal mounting frameworks.

Loop rheostats, when required, may be mounted adjacent to the channel terminals or in a loop pad bay along with other loop rheostats that may be associated with electronic loop repeaters. The latter arrange-

ment concentrates the heat dissipated by these rheostats at a place where it will not be harmful.

Subscriber Set

A channel terminal may also be mounted in a station set box appearing in the knee-well of a subscriber's teletypewriter table. This, called a 130B1 teletypewriter subscriber set, is illustrated in Fig. 10. It contains a line or hybrid coil and balancing network, as well as local circuit resistors and other miscellaneous apparatus. When so mounted, the

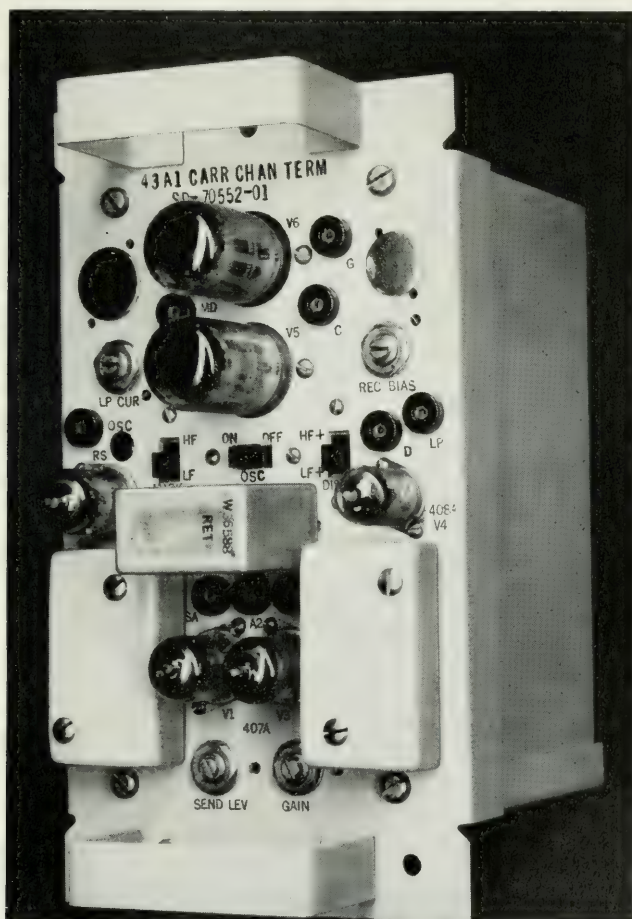


Fig. 7—Channel terminal, front view.

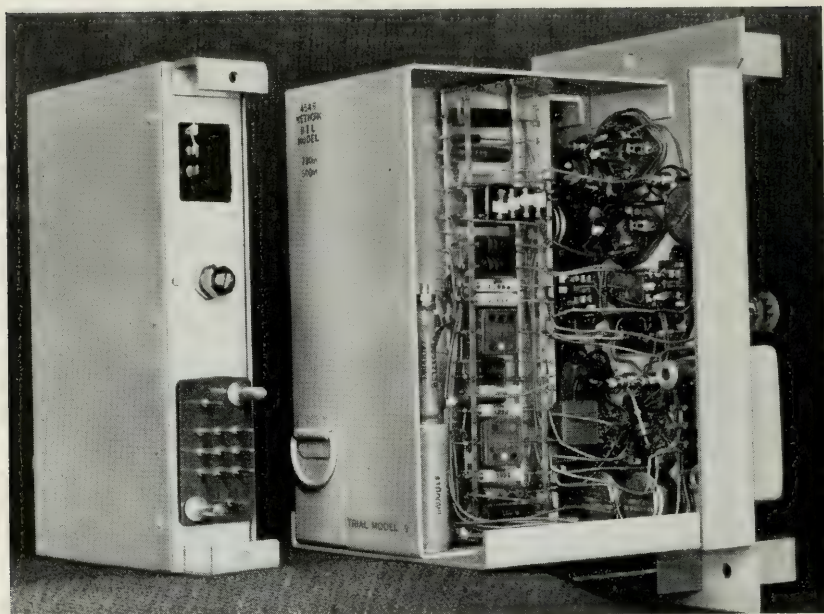


Fig. 8—Channel terminal, rear view, sending network removed.

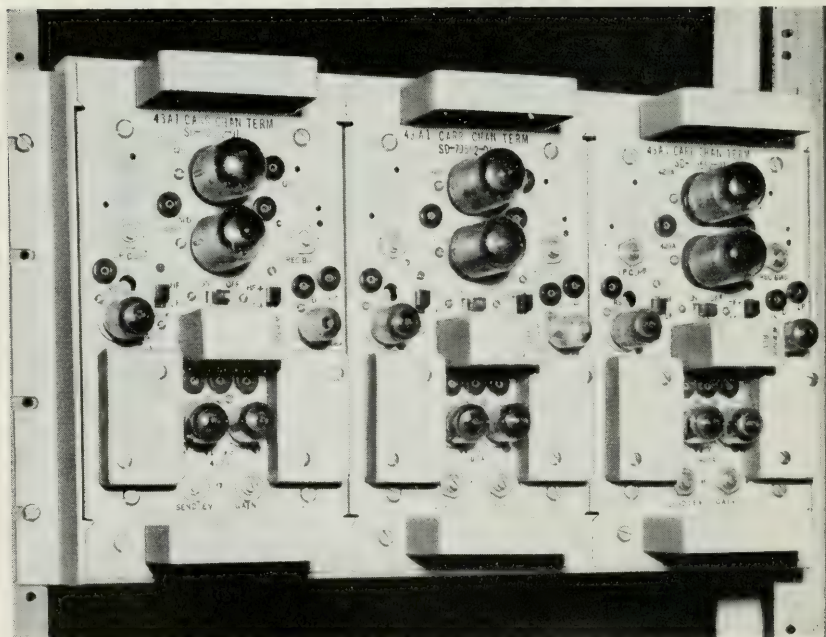


Fig. 9—Three channel terminals mounted on relay rack.

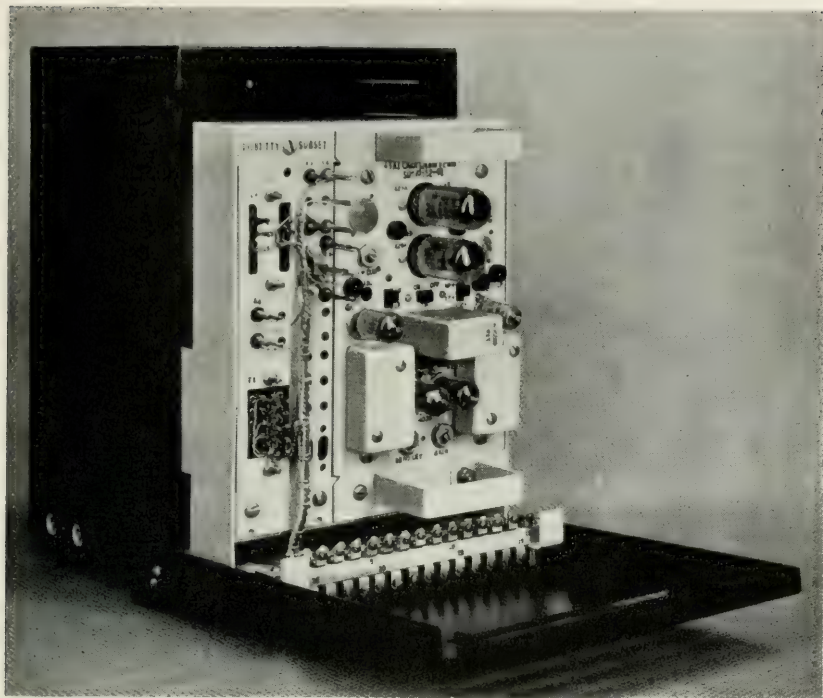


Fig. 10—130B1 teletypewriter subscriber set including channel terminal.

channel terminal is powered by the teletypewriter rectifier, which furnishes 130-volt dc and 20-volt ac power.

The 130B1 set may be employed in private-line or TWX service. In the latter, the application and removal of oscillator plate battery is controlled as described above by the teletypewriter power switch, so that the equivalent of telephone "switch-hook" supervision is attained.

Supervision and transmission are largely independent. The telegraph receiving circuit at the central office terminal remains marking during recall and disconnect signals; hence these supervisory signals do not pass through the cord-circuit repeater to the TWX toll line. Since there is no frequency discrimination in the supervisory receiving circuit, either marking or spacing carrier from the station energizes the supervisory circuit. Hence a communication break (spacing) signal from the subscriber station is transmitted through the operator's cord without any effect on the supervision.

On a TWX call to the subscriber station, a series of alternate marks and spaces, generated by applying 20-cycle ringing voltage to the grid

of the sending tube at the central office terminal, actuates the station ringer, which is connected to the local loop whenever the teletypewriter power switch is in the OFF position.

The circuit which terminates the TWX toll subscriber line at the switchboard office is operable with all existing types of TWX cord circuit repeater. All the features of TWX service, including unattended service, are therefore available.

POWER DRAINS

A channel terminal dissipates about 25 watts. Tube heaters consume about half an ampere at 24 volts and the remainder of the channel terminal, exclusive of its loop-terminating portion, consumes 50 ma at 130 volts. The loop terminating portion dissipates 20, 30 or 62.5 ma at 80 volts, depending upon the type of local circuit employed.

LINE LEVELS

The 43A1 system is capable of working with a great variety of line levels. The send level may be adjusted for any value from +6 dbm downward. The receiving equipment operates satisfactorily with -45 dbm or even -50 dbm. But the levels actually used are controlled by crosstalk and noise conditions in the line.

Receiving levels are normally limited by lightning interference on open wire and by noise on cable circuits. The minimum tolerable levels are about -40 dbm on open wire, -45 dbm on four-wire cable circuits and -35 dbm on two-wire cable.

In Fig. 11, a comparison is made of the effects of static on the 43A1 system and on the 40C amplitude-modulation system. It gives the results of simultaneous tests on the 2465-cycle bands of the two systems, using the static from a record made at Madison, Florida. The 43A1 channel could tolerate about 4 db stronger static than the 40C.

SYSTEM LAYOUTS

A typical circuit layout of the 43A1 system working in the frequency band between the voice and type-C carrier on an open-wire line is shown in Fig. 12. The telegraph circuits extend from 43A1 channel terminals located in a central office, at the left, to 130B1 sets in subscriber stations, at the right. In the central office, the send and receive paths of the channel terminal are combined in a hybrid coil. With the moderate degree of balance provided by the network of this hybrid coil, the allowable difference between send and receive levels of the middle channel may be

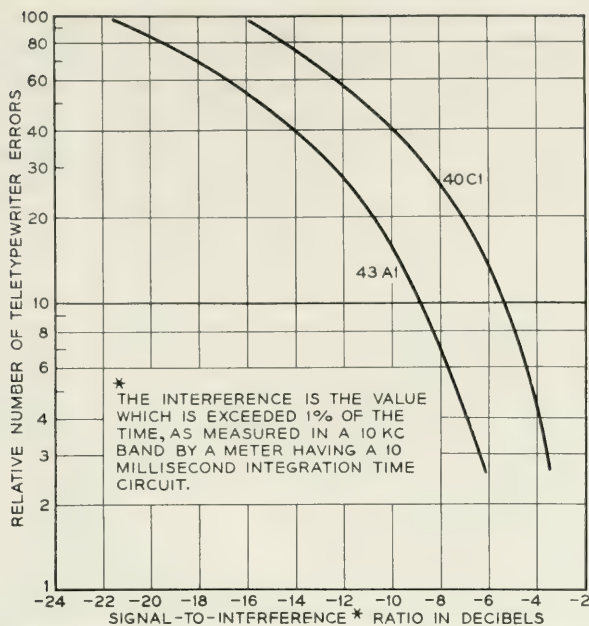


Fig. 11—Comparison of amplitude and frequency shift modulation with static interference.

35 db or more. The telegraph channels are next combined with the voice frequency circuit by means of a 150A filter, and are connected to the composite set and line through the low-pass section of the 121A (type -C) carrier line filter. As a result of the cut-offs of the 150A high-pass and 121A low-pass filter components, the pass band of the telegraph is about 3.7 to 5.4kc. At the outlying terminal of the open-wire line, the telegraph is separated from the voice and type-C carrier circuits by similar filters and connected to the individual subscriber stations by a branching network and branch lines.

The typical arrangement on a two-wire circuit in the voice frequency range is shown in Fig. 13. Six channels are available, using six of the twelve frequency bands for transmission east to west and the other six bands west to east. As in the high frequency case, a branching network and branch lines at the outlying end connect the circuits to the subscribers. Fig. 14 shows a layout in which branch lines are connected at intermediate points in the telephone circuit. At these intermediate points the impedance of the branching network is made high, in order to keep the balance at the telephone repeater from being harmed excessively. Though the network attenuates greatly the signals through it, the telegraph level is usually sufficiently high so that this loss can be tolerated.

The branching network at the outlying terminal has low impedance. Taps on the transformers in the network permit the impedance ratios to be adjusted to suit the line impedances between which the network operates. Since several circuits may pass through this network, a short-circuit on one branch should not be capable of degrading transmission in the other branches. To prevent this, resistances are inserted in series with each branch of such a value that a short circuit will not cause more than 3 db excess loss in other channels. It has been shown by tests that,

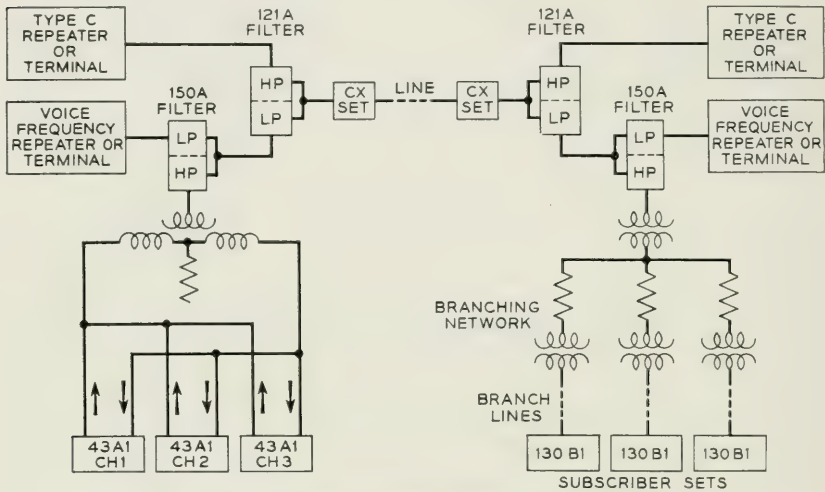


Fig. 12—System layout, above the voice.

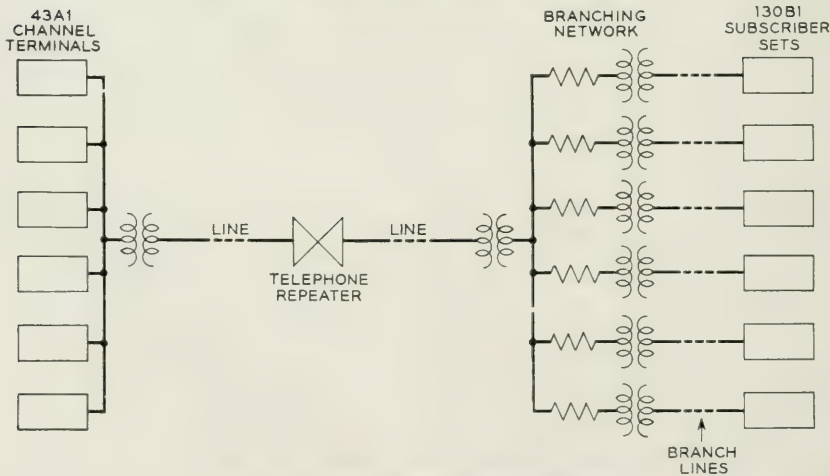


Fig. 13—System layout, in voice band.

in this frequency-modulation system, a sudden loss of 3 db causes little distortion. The series resistances may serve another purpose besides protecting against short circuits. If one branch has a much greater attenuation than the others, the resistance values in series with the shorter branches may be increased so that more energy is directed into the longer branch.

Emergency Circuits

If a circuit containing no intermediate branches fails, a regular message circuit can be patched in to replace it until the trouble is cleared. Fig. 15 shows trunks to be used for making this patch in the case of two-wire circuits. They contain 3 db pads which reduce the signal level to compensate for the change from 0 db transmission level on the regular line to +3 db level on the message circuit.

The 43A1 system may operate also over a four-wire circuit, which accommodates twelve telegraph channels. A patch to an emergency message circuit would then be made at the four-wire patch bay. Since the circuit used for telegraph would usually be similar to those used for telephone message service, no pads to adjust levels would be required for this

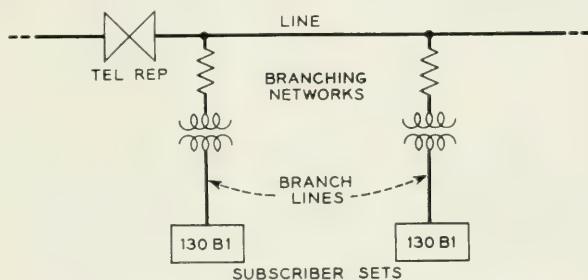


Fig. 14—Intermediate branch lines.

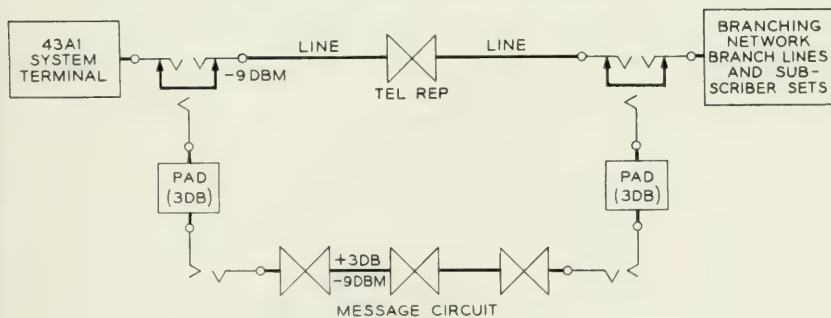


Fig. 15—Emergency circuit.

patch. Obviously echo suppressors must be disabled when a message circuit is used for telegraph.

When the telegraph circuit contains one or more branches at intermediate points, it would be difficult and often impractical to use an ordinary message circuit to replace the telegraph stem in emergencies. The branching location frequently will not be manned and so no one will be available to patch the branch line to the message circuit. In such cases each intermediate branch circuit may be made good over a separate message circuit which is individual to it. Fig. 16 shows this arrangement. A patch trunk is provided between the 43A1 channel terminal at the central office and the telephone switchboard. At the switchboard which is nearest to the intermediate branch subscriber, the branch line is

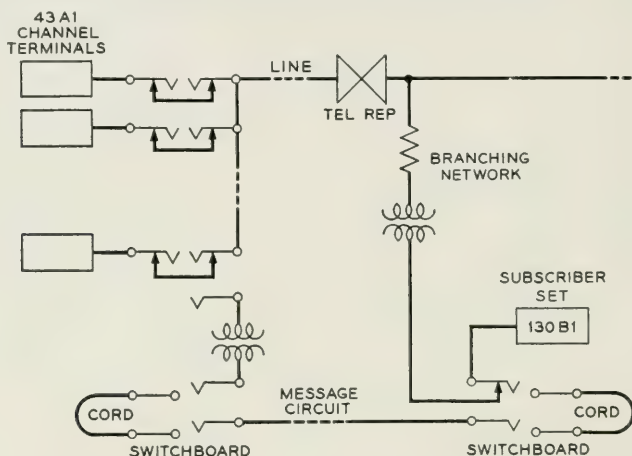


Fig. 16—Emergency circuit for intermediate branch.

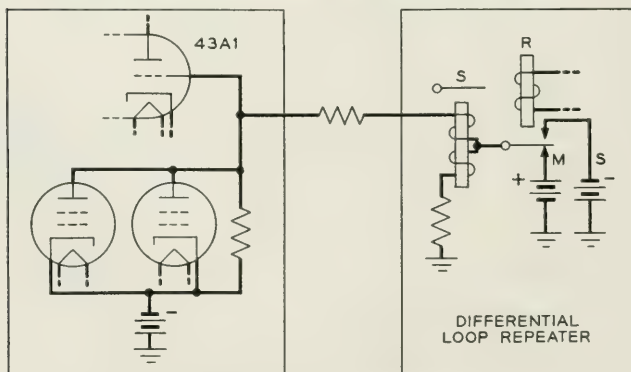


Fig. 17—Connection to other telegraph repeaters.

carried through a cut-off jack. The toll operators then can patch the circuit to the subscriber, thus by-passing the main line when it is in trouble. Since the telegraph energy from only one channel is impressed on the emergency circuit, no adjustment of levels is required.

The channel terminals which are at the central office, shown at the left of Figs. 12 to 16, may be connected to subscribers over dc loops or they may be connected to other types of telegraph repeaters. Fig. 17 indicates the latter connection schematically. Since the loop circuit must supply positive potential to the 43A1 channel terminal, the connecting repeater must be arranged to supply positive battery for marking signals.

FUTURE EXTENSIONS

It is expected that the field of application of the 43A1 system will be broadened by further development over the next few years. More frequencies will be provided, both in and above the voice band. Means will be designed for passing TWX supervisory signals over a direct-current loop from a subscriber station to a channel terminal installed in a nearby central office. The built-in supervisory arrangements of the 43A1 equipment will be exploited to obtain inexpensive straightforward trunks for use both between TWX switchboards and from switchboards to Line Concentrating Units. The supervisory feature will also be employed in private line service to provide an open circuit alarm.

REFERENCES

1. R. S. Caruthers, H. R. Huntley, W. E. Kahl, L. Pedersen, "A New Telephone Carrier System for Medium-Haul Circuits," *Elec. Eng.*, **70**, pp. 692-697, Aug. 1951.
2. B. P. Hamilton, "Carrier Telegraphy in the Bell System," *Bell Labs. Record*, **26**, pp. 58-62, Feb. 1948.
3. J. A. Duncan, R. D. Parker and R. E. Pierce, "Telegraphy in the Bell System," *A.I.E.E. Transactions*, **63**, pp. 1032-1044, 1944.
4. B. Ostendorf, Jr., "New Electronic Telegraph Regenerative Repeater," *Elec. Eng.*, **69**, pp. 237-240, March, 1950.
5. R. L. Case, "Transmission Features of V3 Repeaters," *Bell Labs. Record*, **27**, pp. 94-95, March, 1949.

The Type-O Carrier System

BY PAUL G. EDWARDS AND L. R. MONTFORT

(Manuscript received June 11, 1952)

INTRODUCTION

While the sight of an open-wire toll line is a rarity in many parts of the East, considerable use is made of open-wire facilities in other sections of the country to provide toll and exchange service. At the present time there are about 170,000-route-miles of open-wire in the Bell System which carry some 1,400,000 pair-miles of wire used for toll service. It is estimated that about 60 per cent of this pair-mileage is used for carrier, although only about 10 per cent carries the full fifteen carrier channels, which is possible by employing type-C and type-J carrier systems. It is obvious that some of the remaining line pairs are available for additional carrier growth, provided, of course, the demand for additional circuits exists, and there are carrier systems which can meet these demands economically. Type O is a multi-channel, open-wire carrier system which has been designed to provide, economically, additional circuits in the range from a minimum of about 15 up to a maximum of 150 miles, or more. The type-O system is the open-wire counterpart of the type N short-haul cable system.

Present open-wire toll lines vary from a single-arm line, with one or two pairs of wires, up to lines with five or six arms carrying thirty pairs. These lines may carry long-haul toll circuits up to about 1000 miles in length, short-haul toll circuits up to 150 miles, as well as tributary trunks and exchange circuits. Growth in the past of toll and tributary circuits on these lines has been provided by the addition of single-channel D or H systems, three-channel C systems, twelve-channel J systems or by other similar carrier systems.

The full development of a line for open-wire carrier has, in the past, required expensive line rearrangements. For instance, most lines reach terminal and repeater offices over entrance cables which may be several miles in length. Impedance matching at the junction of the open-wire and cable is required, and is provided by loading the cable at both voice and carrier frequencies, by employing junction line filters using non-

loaded carrier pairs, or by the use of autotransformers. In addition, transposition schemes are needed to reduce the crosstalk coupling between open-wire pairs to tolerable amounts, and longitudinal and metallic filtering is necessary at repeater points to control interaction crosstalk. The C and J carriers have been designed essentially for long-haul use, and when the line transposition costs are added, these systems are likely to be more expensive than adding wire for providing relief for the shorter circuits, which are required in increasing number as the length decreases.

In contrast to the heavy back-bone toll lead carrying both long and short haul circuits is the one or two arm line which may be a secondary route, or a line which terminates in a small town. The demands along this line are for short-haul toll service, trunks between tributary offices and their toll centers, and for exchange service. Because growth has been relatively slow, carrier has been employed only to a limited extent. Single-channel D or H systems may be found on these lines, or an adaptation of the M system for toll service, and possibly other miscellaneous systems. Only minor rearrangements of the line and entrance cables has been necessary because of the small percentage of circuits equipped with carrier facilities.

A typical need for expansion on this type of line occurs when a manual tributary office is cut over to dial operation, and the operators are moved to the toll center. Additional circuits are immediately needed from the toll center to the tributary office because of certain factors introduced when the operators are moved some miles away. Experience has shown the desirability of being able to reach an operator a fairly large percentage of the time because of special services, such as directory information, reports on the availability of toll circuits, service complaints, etc. This requires a substantial increase in the number of circuits to the tributary office in such instances. Development of this line, then, proceeds by adding single-channel carrier systems, until a point is reached where it is necessary to string more copper wire, which is costly and may be in short supply, or to add multi-channel carrier systems.

The situation in Iowa is typical of many areas in the Southern and Western parts of the country. Fig. 1 shows the principal open-wire and cable routes in Iowa used by the Bell System for toll business. The type-K transcontinental cable crosses the state, passing through Davenport and Des Moines on its way to Omaha. Small branch cables serve Muscatine, Cedar Rapids and Atlantic, where the circuits are extended by open-wire facilities. In general, the transcontinental TD-2 radio relay system follows the K carrier route. A coaxial cable route extends north from Des Moines to Minneapolis, connecting at Iowa Falls with short

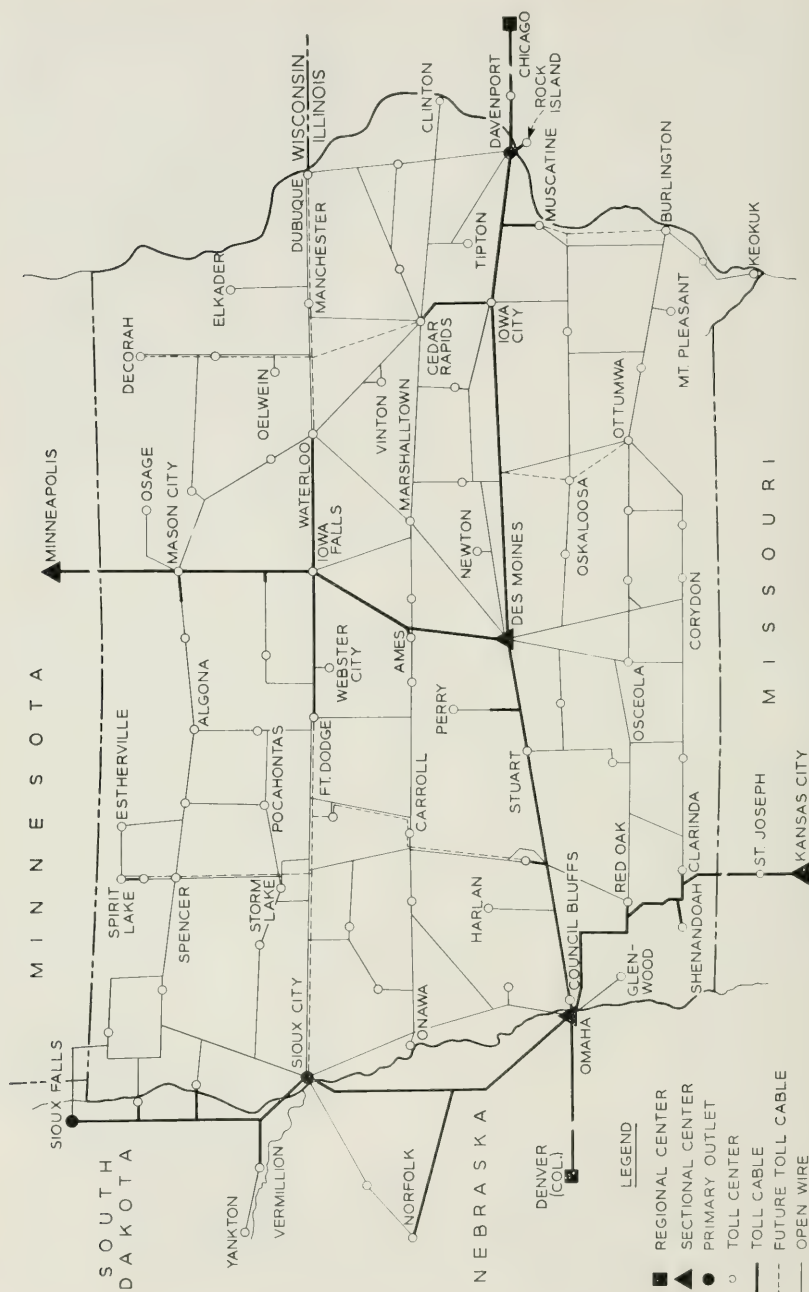


Fig. 1—Principal open-wire and cable routes in Iowa.

K cables to Fort Dodge and Waterloo. Eventually a second cable route will extend across the state through Waterloo and Fort Dodge, as shown by the dashed lines. A second coaxial route cuts across the southwestern corner on its way to Kansas City and a third coaxial route connects Sioux City with Omaha. The rest of the state is served by open-wire lines.

Fig. 2 shows the distribution with length of the Bell System open-wire short-haul toll and tributary circuits in Iowa in 1950, including both voice-frequency and carrier facilities. It will be noted that 95 per cent of the circuits are less than 100 miles in length, while 90 per cent are less than 70 miles, the point where type C systems just become economical. For tributary circuits 98 per cent are less than 30 miles in length. There are a total of some 2700 toll and tributary circuits in Iowa. In addition, there is also a sizable connecting company development. Fig. 3 is a distribution of the number of circuits per group, where a group is composed of the circuits used for via or terminating business between two towns. There are about four circuits per group for short-haul toll and two circuits per group for tributary service in the median case. As the dial conversion program proceeds the average number of circuits per tributary group is expected to increase.

Because of the short distances involved, and the small number of circuits per group, carrier development in Iowa has been restricted, to a large extent, to single-channel systems, and type M. Only four or five M channels can be operated on a given open-wire line, and while these systems have some transmission disadvantages, they have been employed to a large extent. However, further M development is blocked because of the expense of isolating M systems from adjacent lines. There are a few C systems on such routes as Sioux City-Spencer-Mason City, Waterloo-Dubuque, Muscatine-Keokuk, and Atlantic-Spencer.

The type-O system, therefore, is being made available to provide short-haul toll and tributary circuits on open-wire lines in the range from 15 to 150 miles. When completely developed it will provide four four-channel systems in the frequency range from 2 to 156 kc, as shown on Fig. 4. The use for separate channels of both sidebands of a single carrier, called a "twin-channel," results in economical use of the frequency space. The four channel systems are designated OA, OB, OC, and OD respectively, and cover substantially the same frequency range as the C and J systems.

Considerable attention has been given to keeping the line rearrangements as simple as possible. The use of non-loaded entrance cable is proposed, and simple arrangements are available for adding O groups

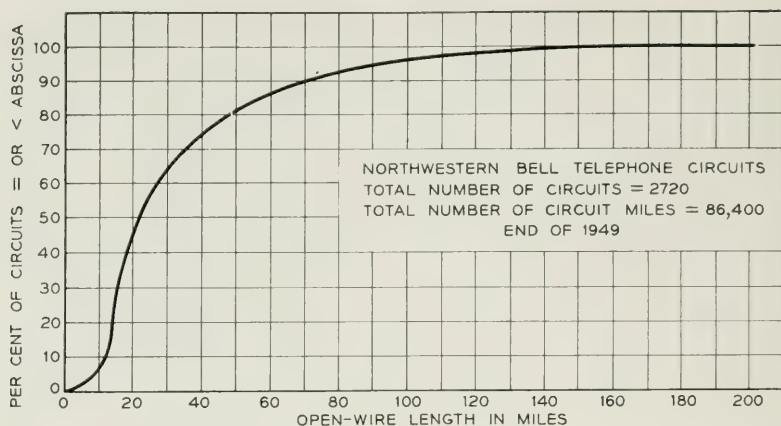


Fig. 2—Distribution of circuit lengths in Iowa in 1950.

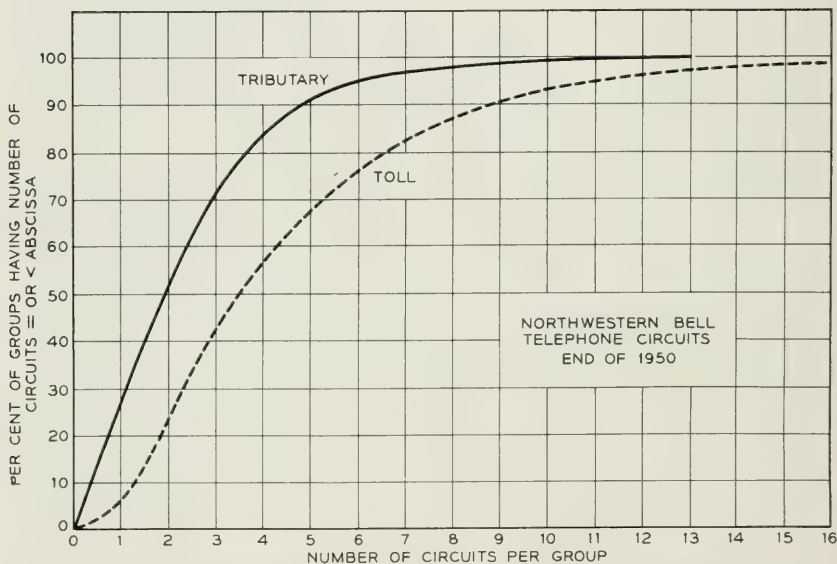


Fig. 3—Distribution of circuits per group in Iowa in 1950.

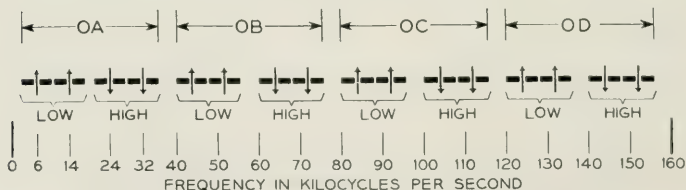


Fig. 4—O carrier telephone—frequency allocation.

above existing carrier systems, such as C, which has a top frequency of 30 kc. With the aid of the compandor, it is possible to apply the OB system to practically any open-wire pair transposed for C carrier operation, thus nearly doubling the number of circuits without additional line rearrangements. Transposition arrangements which are expected to be less expensive to apply are being made available where the higher-frequency type-O groups are involved. Line losses of the order of 35 to 40 db can be spanned under normal wet weather conditions, and 50 db loss under sleet conditions with some transmission impairment, since sleet is relatively infrequent in occurrence. This will result in repeater spacings of the order of 50 miles in sleet areas for the OB group, and 100 miles in other areas. For the higher frequency groups (OC and OD) these spacings will be approximately halved.

CHOICE OF DEVELOPMENT APPROACH

The type-O carrier system followed the type-N system closely in time, and, in effect, covers the same range of circuit lengths for open wire lines that N provides for cables. It was both natural and expedient that many of the N features were carried over directly into the O design. It was necessary, however, to make important distinctions as well. These similarities and differences will be discussed in some detail.

The transmitting and receiving voice frequency subassemblies are reused with substantially no modification. This provides the O system with the same compandor and the same 3700 cycle signaling system as used in N.

An important difference between the two systems is concerned with the use of single sideband in the O rather than double sideband as in the N. This choice is an economic one. The double sideband system is relatively easier to design and less expensive than the single sideband arrangement. The use of double sideband in cables is practicable in many cables because of the relative abundance of conductors as compared with open wire pairs. In some cases the use of single sideband in cables may be attractive as compared with the cost of new outside plant for certain length ranges.

Another distinction between the two systems is the provision of circuits in smaller groups in O. In N the basic group is 12, although systems may in some instances be partially equipped. In O, the desire to furnish smaller circuits groups resulted in the choice of a basic four-channel group. The full complement per pair for O, including a channel replacing the voice circuit, is sixteen channels.

The regulation problem is more severe in the O design. It is necessary

to provide sufficient regulator range to accommodate line variations due to wet or icy lines. The repeater and receiving group regulator range common to four channels is in the order of 40 to 50 db, or approximately four times the regulating range of an N repeater. The range of the twin-channel regulator is comparable to the N individual channel regulator, but the O regulator is shared by two channels forming adjacent sidebands of the same carrier.

The use of a single sideband imposes more severe requirements on channel band filters. The use of a material called ferrite, in combination with a crystal, affords an efficient channel band filter in a small space when compared to previous single sideband channel filters employing air-core coils. Ferrite coils are employed in a coil-condenser type of filter to provide separation for the various four-channel groups. While the N system employs only receiving channel band filters, O has filters in both the transmitting and receiving terminals.

The O system employs the double modulation principle for all groups. This arrangement permits the use of only four channel band filter designs for all of the 32 channel frequency allocations. The frequency range for these basic channel bands has been selected to provide the most economical overall filter design.

The use of die-castings has been extended in a number of ways. Notable among these is a die-cast framework, used in both the terminal and the repeater. The plug-in technique has been expanded to provide plug-in filters for channel and group band filters.

DESCRIPTION OF THE SYSTEM

The system will be described first by block schematics, second, by transmission characteristics, and finally by photographs. This description will show representative features rather than describe the system completely.

While the description will cover the complete O plan, it should be pointed out that the OB system is the first to be made available. It will be followed by the other O systems.

In the schematic description, where a unit is common to all systems the designation is "Type O." Where the arrangement is different for the several systems, a particular designation is used, such as "Type OB," etc.

Schematics

The O modulation plan is shown on Fig. 5. The single-sideband channel filters for all groups are in the frequency range from 180 to 196 kc.

By the use of different group carrier frequencies the several four-channel groups are placed in their various locations. As indicated by Figs. 4 and 5, high and low group assignments are used for the two directions of each four-channel system. A repeater is provided for each four-channel system and, except in the case of OA, the high and low frequency groups are "frogged" at each repeater, as in the N system.

Figure 6 shows a block diagram of a complete O system. On this figure, and in general on other figures showing filters, a letter code is assigned to designate the kind of filter, with a subscript letter to indicate the particular system in which the filter is used. The several filter designations and number codes are collected for reference in Table I. Much of the apparent complexity of the system, particularly as regards filters, arises out of the use of a single pair for both directions of transmission. Another complexity is occasioned by the requirement that a complete complement of O systems will not always be provided. For example, OB, OC, and OD systems may be used above an existing C system, or similar systems.

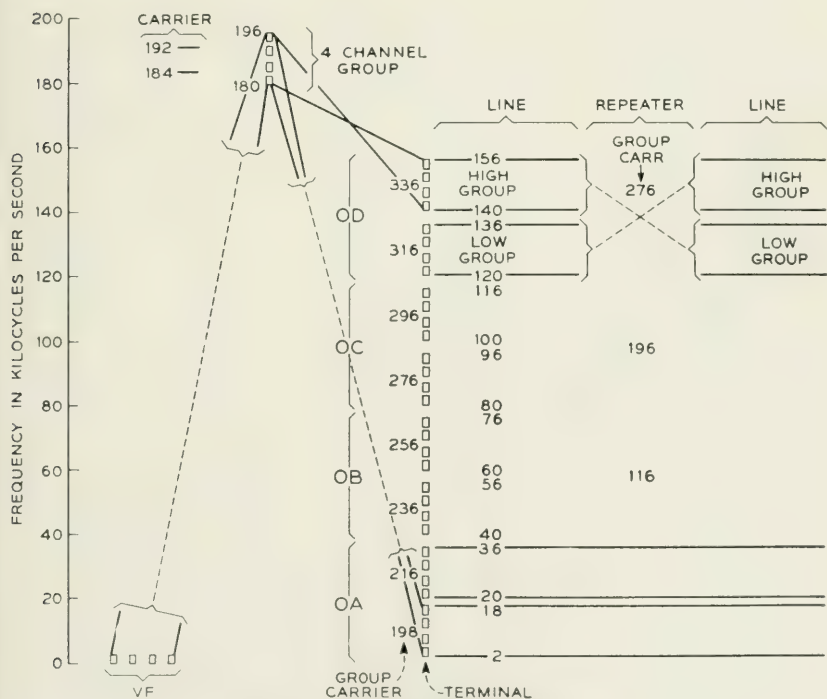


Fig. 5—Type-O modulation plan.

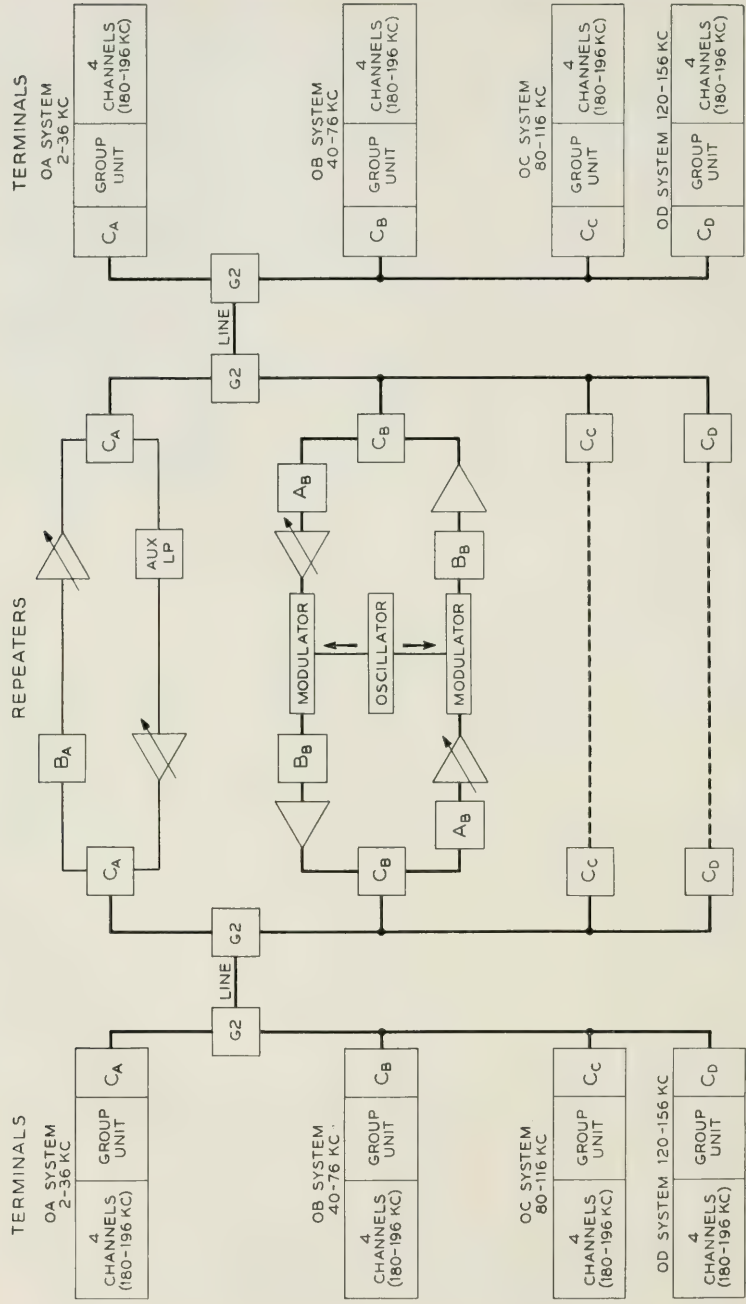


Fig. 6—Type-O carrier system.

Line filters (G) are provided to separate the OB, OC, and OD Systems from the OA frequency ranges. Because of the lower attenuation and slope in the OA frequency range, and the better line coupling factors, the repeaters do not "frog" the high and low groups.

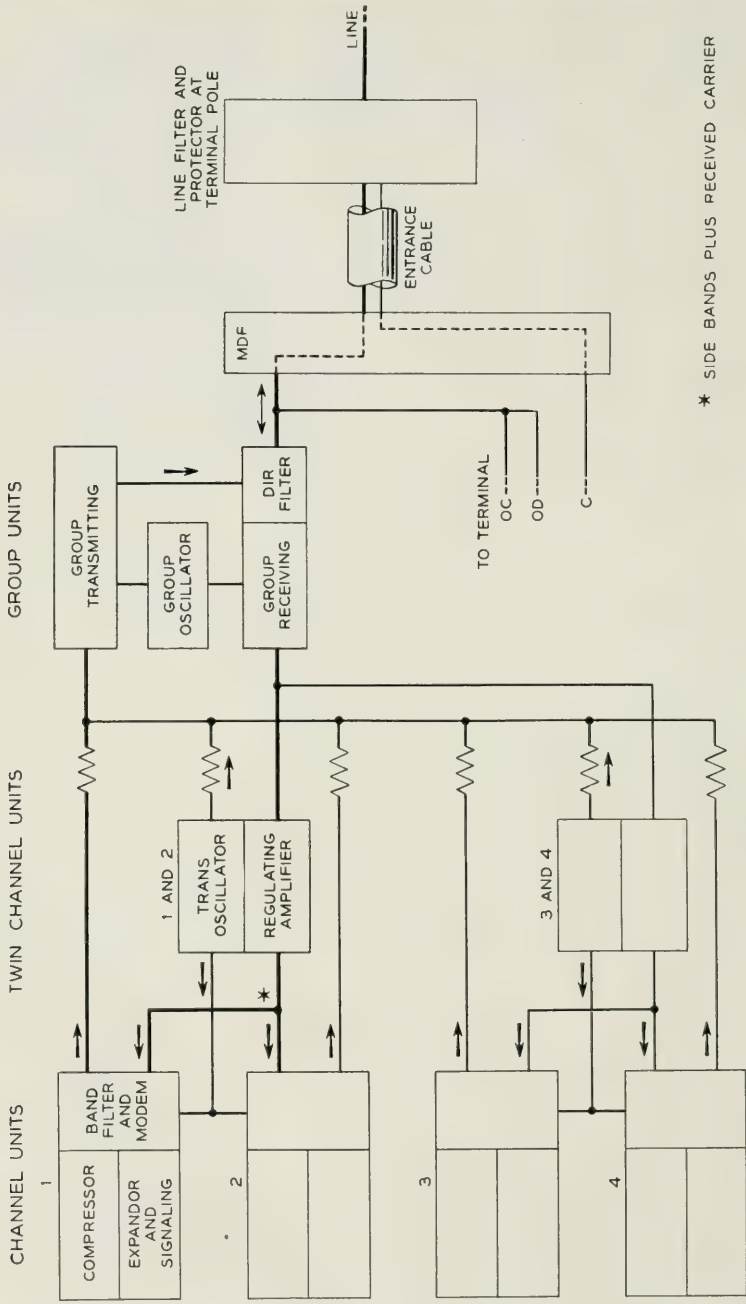
The OB Carrier Terminal

A block diagram of a typical carrier terminal is shown on Fig. 7, in this case the OB terminal. The terminal is comprised of four channel units, two twin-channel units, group transmitting and receiving units, and a group oscillator. An oscillator in each of the twin channel units supplies carrier to the transmitting channel modulators. The same oscillator supplies transmitted carrier for two associated sidebands. The original carrier is balanced out in the transmitting modulator. This method results in a more accurate control of the transmitted carrier level. The group oscillators supply the necessary frequencies to the

TABLE I

Location and Function	Filter Symbol	Filter Codes for Each System				
		Common	OA	OB	OC	OD
TERMINAL ONLY						
Transmitting low pass.....	None	168F				
Receiving low pass.....	None	169G				
Carrier pickoff (184kc).....	F1	532A				
Carrier pickoff (192kc).....	F2	532B				
Signal pickoff.....	None	169A				
Channel band pass.....	None	529A				
Channel band pass.....	None	529B				
Group transmitting.....	E	540A*				
Group receiving.....	A + D		530J	531B	530C	530F
Group receiving.....	B + D		531F	531C	530D	530G
TERMINAL AND REPEATER...						
Directional.....	C		530H	530A	530B	530E
Line.....	G $\begin{cases} \text{G2} \\ \text{G1} \\ \text{G1} \end{cases}$	$\begin{cases} 219\text{S}\dagger \\ 537\text{A}\ddagger \\ 538\text{A} \end{cases}$				
REPEATER ONLY						
Auxiliary.....	A + B			531A	531D	531E
	A		530K			
	B		530L			

* Except OA system. † Cut-apart region between 36 and 40 ke. ‡ Cut-apart region between 30 and 40 ke. 538A is a 537A filter with housing and protectors for pole mounting.



* SIDE BANDS PLUS RECEIVED CARRIER

Fig. 7—Type-OB carrier terminal.

transmitting and receiving group units to translate the sidebands between the frequency range of the channel band filters and the correct line frequency allocation. The group receiving unit contains the directional filter for separating the four-channel transmitting and receiving groups at line frequencies. Multiple points are indicated for the connection of other O carrier systems on the same pair.

Channel Unit

A block diagram of the channel unit is shown on Fig. 8. As indicated by the dashed line, the channel unit is comprised of four parts which are interconnected by plugs and jacks. These are:

1. *The Compressor Sub-Assembly.* This unit contains the compressor and a terminating arrangement to permit the system to be used for four-wire termination, or for two-wire termination at non-gain locations, i.e., those without switching pad control.

2. *The Expander Sub-Assembly.* This unit contains the expander as well as the signal transmitting and receiving equipment.

3. *The Carrier Frequency Sub-Assembly.* This unit contains the transmitting and receiving modulators, and is arranged to receive the plug-in transmitting and receiving channel band filters.

4. *The Transmitting and Receiving Band Filters.* These are combined in a single plug-in unit.

Items 1 and 2 are practically identical to the corresponding sub-assemblies for N carrier. Each channel receives its carrier supply for the transmitting side from an oscillator in the twin channel unit. On the receiving side the carrier is obtained by selecting and amplifying the transmitted carrier.

The frequencies indicated on Fig. 8 are the same for all O systems, and apply to two of the four channels in the group. Figs. 4 and 5 show go and return channels in high and low frequency assignments in the same O system. However, as shown on Fig. 9 covering the OB system, the frequency assignments applying to the channel band filters are above any of the O line frequencies and are in the frequency range from 180 to 196 kc for both transmitting and receiving channel band filters.

Transmitting and receiving channel band filters are paired in a single plug-in unit. In order to reduce the number of kinds of paired channel band filters (transmitting and receiving in the same plug-in unit) the pairing has been done in a special way. If, for example, transmitting assignment 180-184 were paired with receiving assignment 180-184, etc., four different kinds of paired filters would result. Instead, assignment 180-184 is paired with assignment 192-196, and by making the

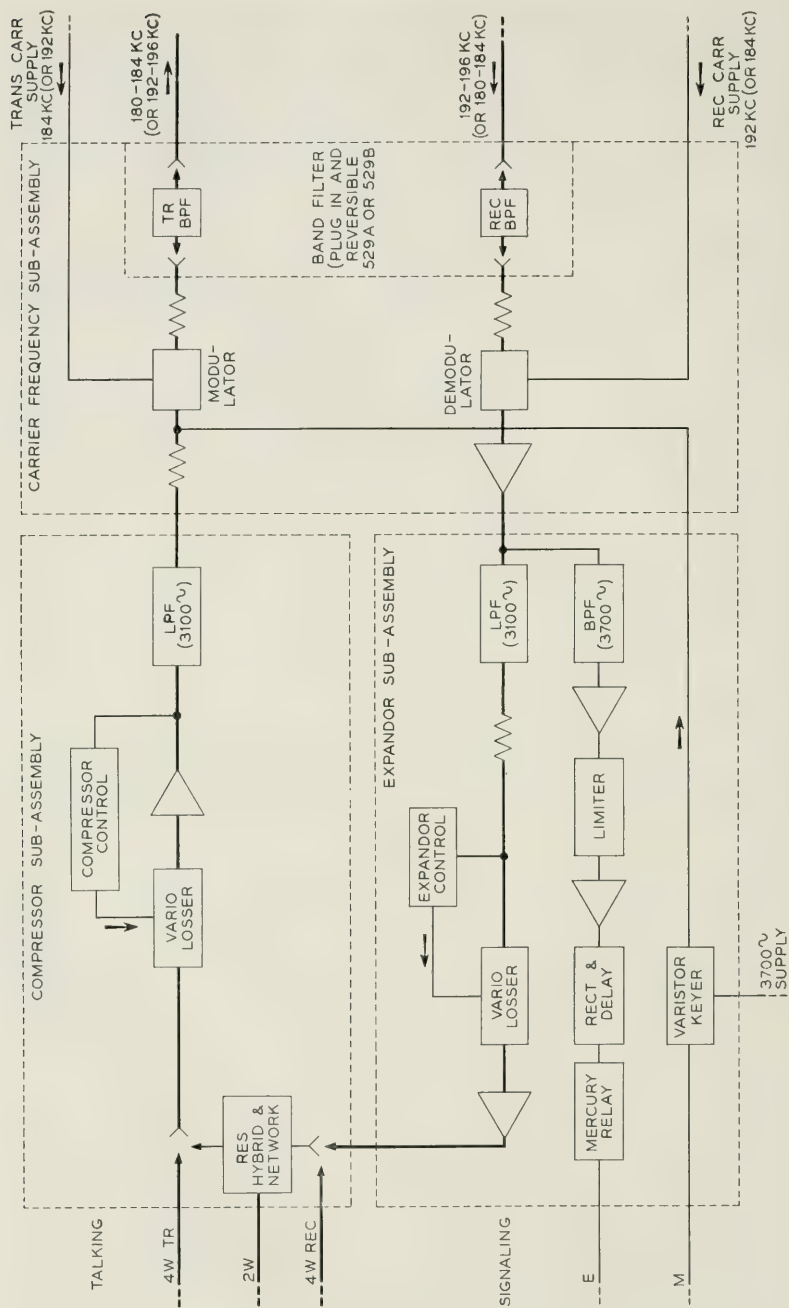


Fig. 8—Type-O channel unit.

plug-in filter reversible in its socket, this grouping can be made to serve two channels as follows:

$$\left\{ \begin{array}{l} 180-184 \text{ Transmitting} \\ 192-196 \text{ Receiving} \end{array} \right\} \text{ and } \left\{ \begin{array}{l} 192-196 \text{ Transmitting} \\ 180-184 \text{ Receiving} \end{array} \right\}$$

A similar paired filter serves

$$\left\{ \begin{array}{l} 184-188 \text{ Transmitting} \\ 188-192 \text{ Receiving} \end{array} \right\} \text{ and } \left\{ \begin{array}{l} 188-192 \text{ Transmitting} \\ 184-188 \text{ Receiving} \end{array} \right\}$$

Thus only two basic kinds of paired channel band filters are required, rather than four kinds. In these filters, as well as the reversible group filters, the filter designations are so arranged that when the filter is in place the proper filter designation is in view.

Twin Channel Unit

The twin channel unit is shown in somewhat more detail in Fig. 10. There are two kinds of twin channel units to serve the four channel assignments, and the frequencies shown on Fig. 10 correspond to those shown on Fig. 8. (and Fig. 9). A transmitting carrier adjustment permits the transmitted carrier level to be set properly in relation to the sideband levels. In order that the group regulators may function primarily on the carrier, and thus be substantially independent of voice or signaling sidebands, the carrier is transmitted approximately 6 db above the sideband level.

On the receiving side of the twin channel unit a regulating amplifier controls the received level of both sidebands. It does this from the carrier

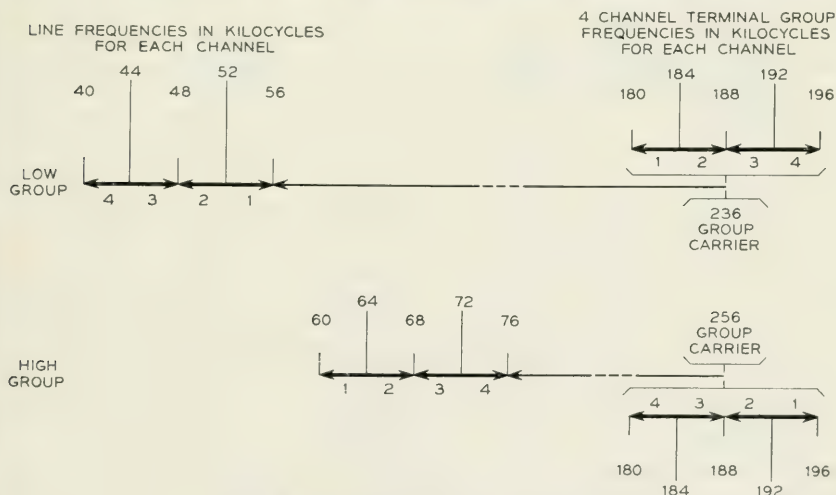


Fig. 9—Type-OB system frequencies.

picked off by a narrow band crystal filter. This same carrier is supplied to the receiving side of the channel units for demodulating the associated sidebands.

Group Transmitting Unit

The OB group transmitting unit is shown on Figure 11. It receives the four sidebands and two transmitted carriers and places them in the proper high or low line frequency assignment. The transmitting group unit, depending on the optional connection to the group oscillator (Fig. 11), can be either a high group transmitting unit or a low group transmitting unit.

For convenience the noise generator is contained in the group transmitting unit. On very quiet circuits this noise source provides a means of masking crosstalk. In ordinary usage the noise thus provided is not noticeable on the circuit, but is sufficient to reduce the chance of hearing intelligible crosstalk to a small value.

Group Receiving Unit

The OB group receiving unit is shown on Fig. 12. It is comprised of an amplifier and a regulator-modulator arrangement equipped with plug-in filters. The same basic arrangement is used for all receiving group units, as well as for all repeaters. Only the plug-in filters, and the frequencies from the associated oscillators are different. The directional

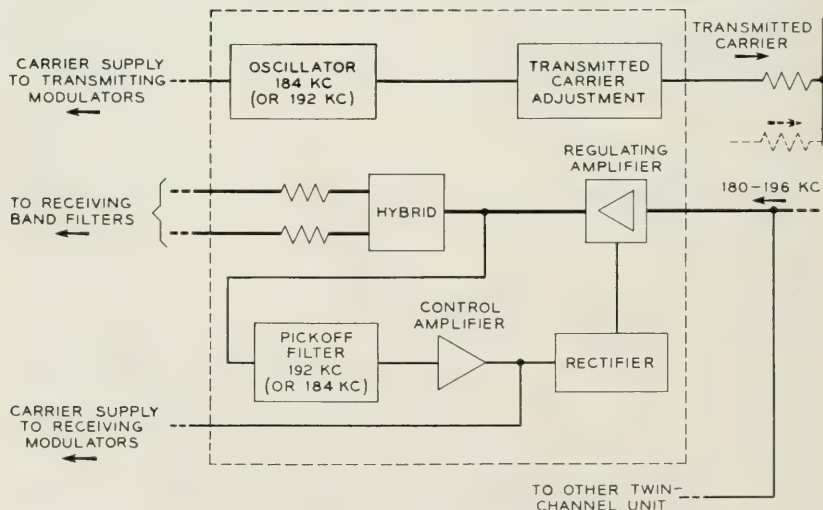


Fig. 10—Type-O twin channel unit.

filter is reversible as well as plug-in and thus serves either high or low groups. The receiving group band filter and its associated auxiliary filter have the same physical arrangement as in the repeater but they are never reversed.

A dc feedback type regulator controls the gain of the regulating amplifier, and operates principally on the two received carriers, although the sidebands are fed back also.

Group Oscillator

The group oscillator contains two oscillators for supplying the group transmitting and group receiving units. These oscillators are interchangeable (by strapping) and permit the group units to operate in either the high or low groups. For convenience the 3700 cycle signaling oscillator, common to all four channels, is contained in the group oscillator.

Repeater

As indicated on Figs. 5 and 6, a repeater is provided for each four-channel system. An amplifier, regulator and modulator arrangement serves each direction. The directional bands are routed through the repeater by directional and auxiliary band pass filters as indicated on Figs. 13 and 14. At each repeater (except OA) the high- and low-frequency groups are "frogged" to improve transmission, particularly as

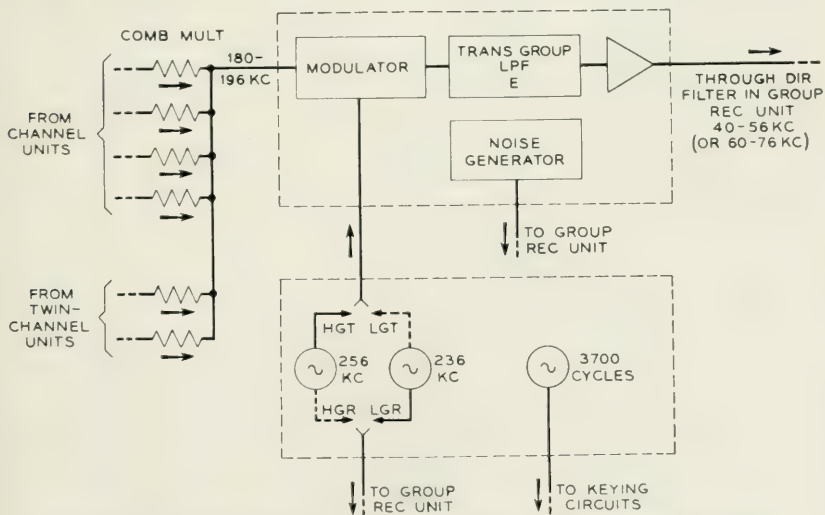


Fig. 11—Type-OB group transmitting unit and group oscillator unit.

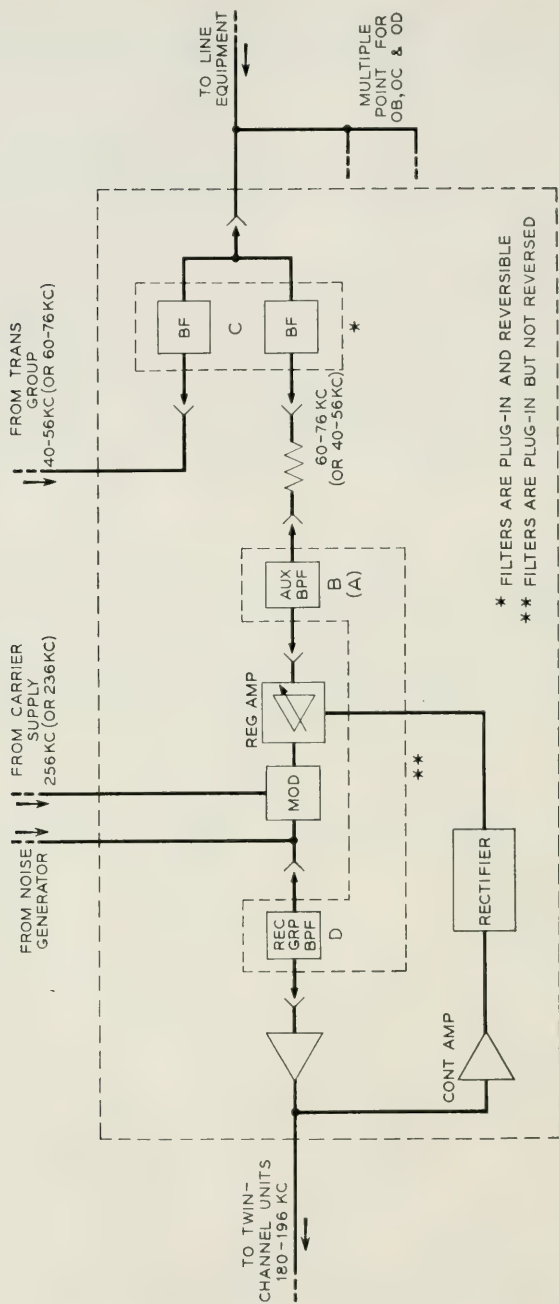
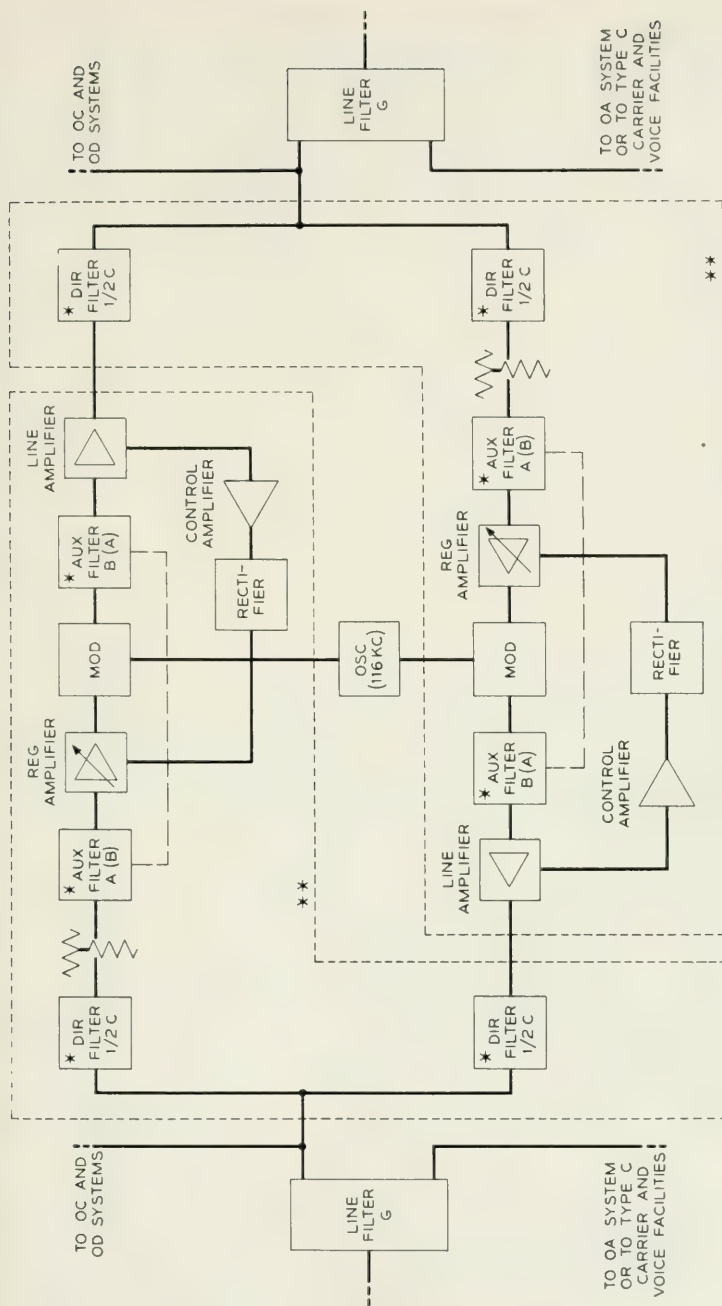


Fig. 12—Type-OB group receiving unit.



★★ BASIC AMPLIFIER USED FOR OB, OC AND OD FOR REPEATERS AND GROUP RECEIVING UNITS

Fig. 13— Type-OB repeater.

regards automatic equalization of attenuation slope with frequency, and to obviate the necessity for additional line treatment.

Some repeaters receive low group frequencies and transmit high group frequencies for both directions. Other repeaters receive high group frequencies and transmit low group frequencies. In N two kinds of repeaters were required. In O the reversal of the filters in their sockets provides both kinds of repeaters, and presents the proper designation to view. A regulator is provided in each direction of the same kind as in the group receiving unit.

It should be noted at this point that the filters internal to the repeater (as opposed to directional filters) differ from the filters used in the receiving group units. This is because the repeater always accommodates line frequencies on both sides of the amplifier, while the receiving group unit accommodates line frequencies on one side (which correspond to the repeater line frequencies) but always must supply channel band filter frequencies at the group amplifier output.

Fig. 14 shows in somewhat greater detail than Fig. 13 the filter arrangement for an OB repeater.

TRANSMISSION CHARACTERISTICS

The overall channel band width is illustrated in Fig. 15. The approximate frequency cutoffs are similar to N but for various reasons the several channels may show somewhat greater differences. The O system, being a single sideband system, has a filter cutoff at low (voice) frequencies, which the N does not have. Differences may exist between

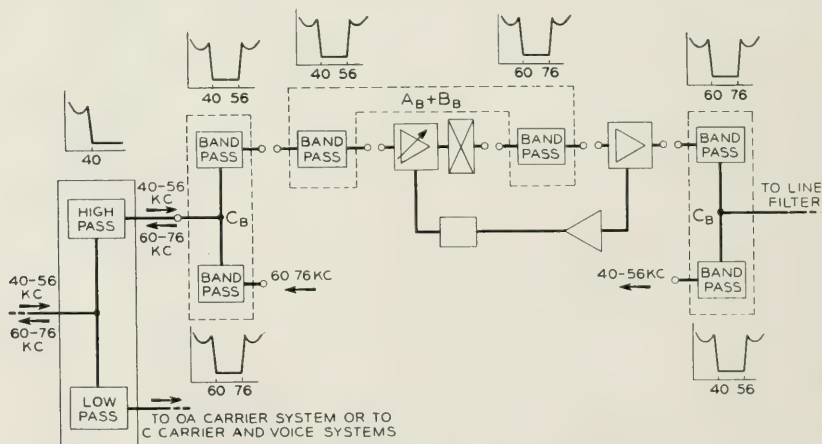


Fig. 14—Type-OB repeater filter arrangement.

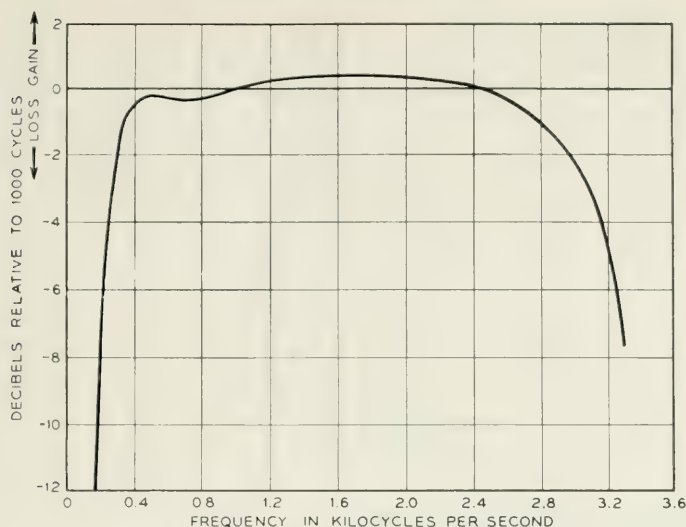


Fig. 15—Net loss frequency characteristic.

upper and lower sideband filters. In addition, O has transmitting band filters while N does not.

A situation is of interest which applies to both N and O, as well as to any other system employing the type of compandor controlled from the voice energy. Different frequency characteristics will be obtained with the compandor operating and with the compandor controls locked. Neither of these necessarily corresponds to the operating condition with speech or music. With the compandor controls free and using single frequency test tone, the characteristic obtained is a combination of the frequency characteristics of the line and control circuits. With the controls locked, the characteristic is that of the line only. If the control circuit is substantially flat, there will be little distinction between the measurements. The curve of Fig. 15 is of the type obtained with free controls and with a substantially flat control circuit.

A typical overall channel load characteristic is shown on Fig. 16. This characteristic includes not only the load curve of various amplifiers, modulators, etc., but shows also the order of match of the compressor and expander load characteristics. This is a match of curves having 2:1 slopes on a db basis over a wide range of volumes.

A typical overall net loss variation for a non-repeated circuit is shown on Fig. 17. Principally because of the line regulator in the group receiving units (Fig. 18) a wide range of line loss is covered. A similar regulator is included in each repeater, and the extension of a system

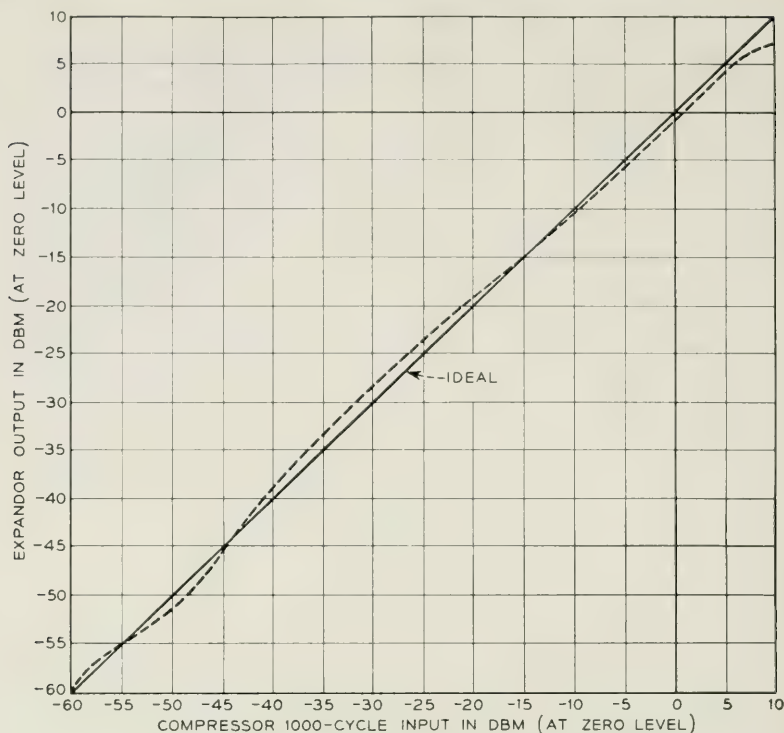


Fig. 16—Typical over-all channel load characteristic.

with repeaters will not result in a substantial change of the net loss variation, assuming the repeater section losses do not exceed the range of the regulators.

The line regulator is assisted by the twin channel regulator, for which a characteristic is shown on Fig. 19. This regulator is similar to the individual channel regulator of N, and serves two channels having a common carrier. This fact alone does not materially change the effectiveness of regulation since the carrier is adjacent to the sideband which

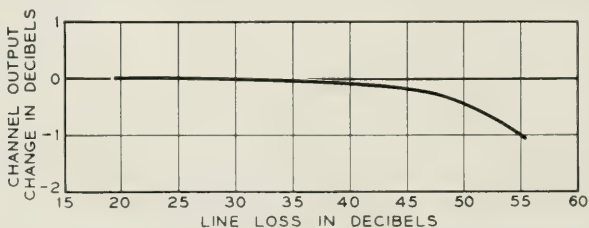


Fig. 17—Typical over-all channel net loss variation, nonrepeated.

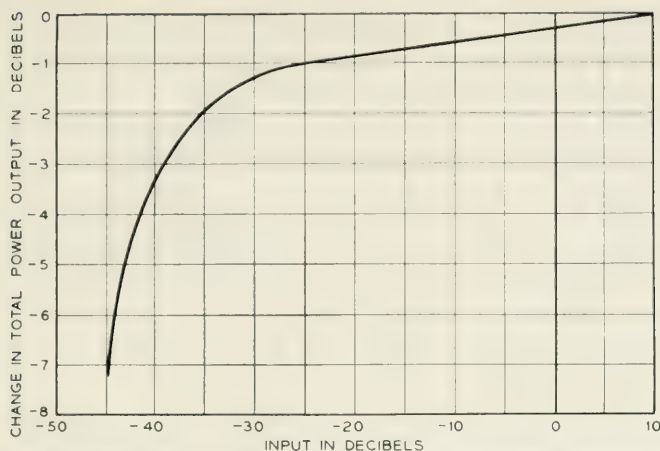


Fig. 18—Typical group receiving and repeater regulator characteristic.

it controls in any case. There are other important differences, however, between N and O channel regulators. In N the channel regulator follows the channel band filter and thus tends to compensate for its flat transmission variations. Also the N regulator is controlled from the demodulator dc output and thus compensates in some degree for demodulator variations. In O, the twin-channel control is ahead of both the channel band filter and the channel demodulator, and therefore does not make up their variations.

A statement might be interpolated at this point to emphasize that the relative advantages of single-sideband and double-sideband transmission are by no means easily listed and evaluated, since the differences are many and devious, some necessarily and some fortuitously. An example worthy of note is that in N it is necessary to be concerned about relative phase shift of the sidebands and in the instances of longer cir-

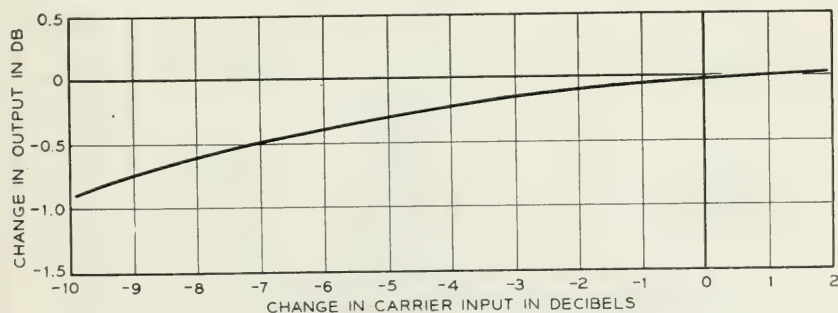


Fig. 19—Typical twin-channel regulator characteristic.

cuits, perhaps to equalize this phase shift in order to prevent serious reduction of signal output, or variation in channel net loss with frequency. No such concern applies to O.

In regard to filter characteristics, it seems obvious that complete coverage is not feasible in this description. Instead typical curves only will be shown.

Fig. 20 shows the general characteristics of filters for separating the wanted sideband from the carrier and unwanted sideband. The transmitting and receiving filters have similar shapes. The carrier pick-off filter characteristic is shown in the same figure. Fig. 21 shows the filter characteristics for separating the voice and signaling (3700 cycle) functions.

Another filter case of interest is the line filter for separating, for example, the OA system from the OB system, and from the OC and OD systems, as well, if they are employed. Fig. 22 shows the configuration and loss characteristics of the G1 (537A or 538A) filter. A C_B directional filter (530A) characteristic is shown on Figure 23. This filter assembly includes two filters to accommodate the OB high and low group assign-

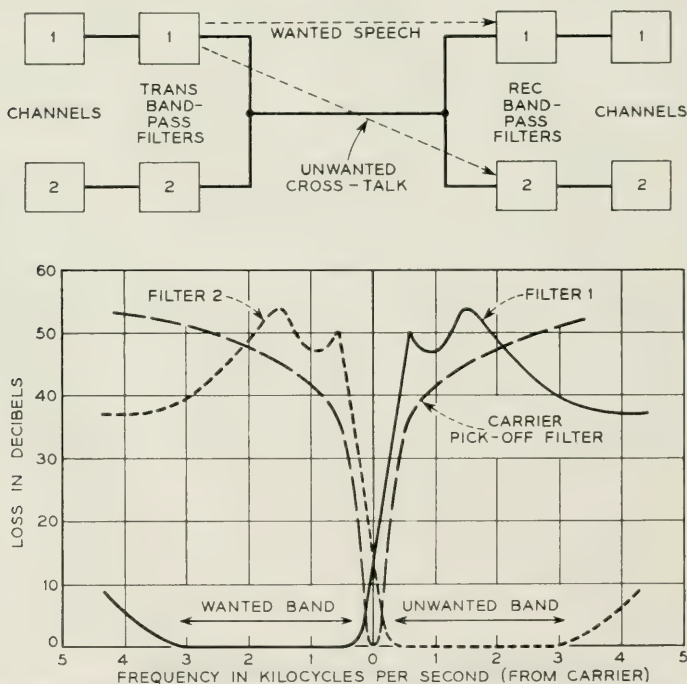


Fig. 20—Typical channel band and carrier pick-off filter characteristic.

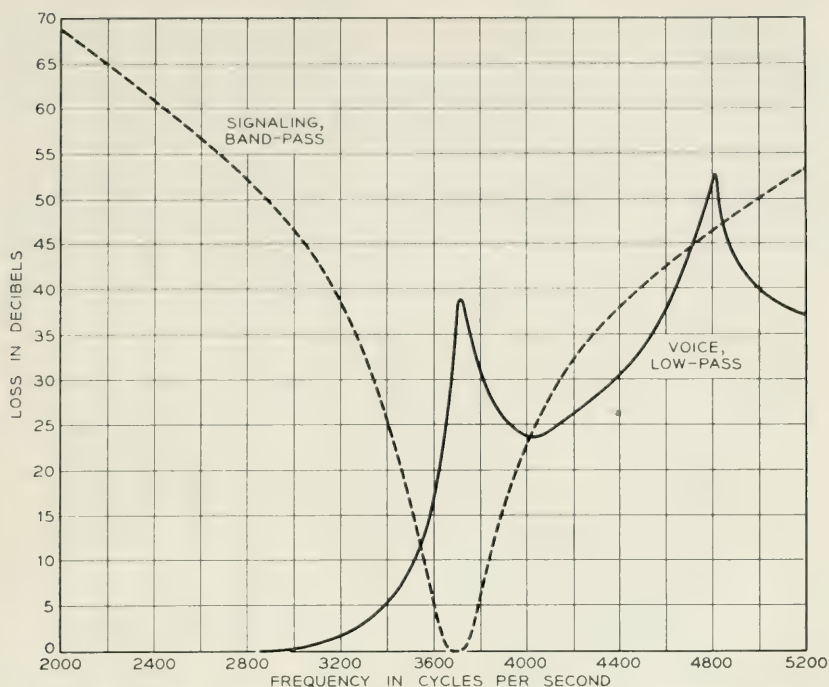


Fig. 21—Typical receiving low pass and signaling filter characteristics.

ments. Similar characteristics apply to the $A_B + B_B$ auxiliary filter (531A). The A and B characteristics are used in the group receiving filters $A_B + D_B$, (531B) and $B_B + D_B$, (531C). The D filter is a band-pass filter with relatively gradual cutoff to pass the 180–196 band for the channel filters, and has peaks at the group carrier frequencies of 236 kc and 256 kc. These filter characteristics are not shown.

PHYSICAL ARRANGEMENTS

A four-channel carrier terminal is shown on Fig. 24. This terminal includes four channel units shown in detail on Figs. 25, 26, 27 and 28. In Fig. 28 the unit is separated into the three sub-assemblies, of which as noted above, the two voice frequency sub-assemblies are substantially the same as for N carrier. The carrier sub-assembly with its plug-in channel band filters is shown on Fig. 29.

The interior arrangement of the plug-in unit containing the transmitting and receiving band filters is shown on Fig. 30. This assembly contains an adjustable ferrite inductance, a miniaturized transformer,

and a crystal with the necessary fixed and adjustable capacitors. The small size is made possible partly by the high Q ferrite coil, and partly by the circuit configuration employing it. As compared with filters employing air-core coils and having comparable cutoffs, the reduction in size of these filters is very striking.

The group receiving unit is shown with its plug-in filters on Fig. 31. The filters are held in place by stud screws and nuts. The arrangement shown on Fig. 31 is also used with different filters for the repeater ampli-

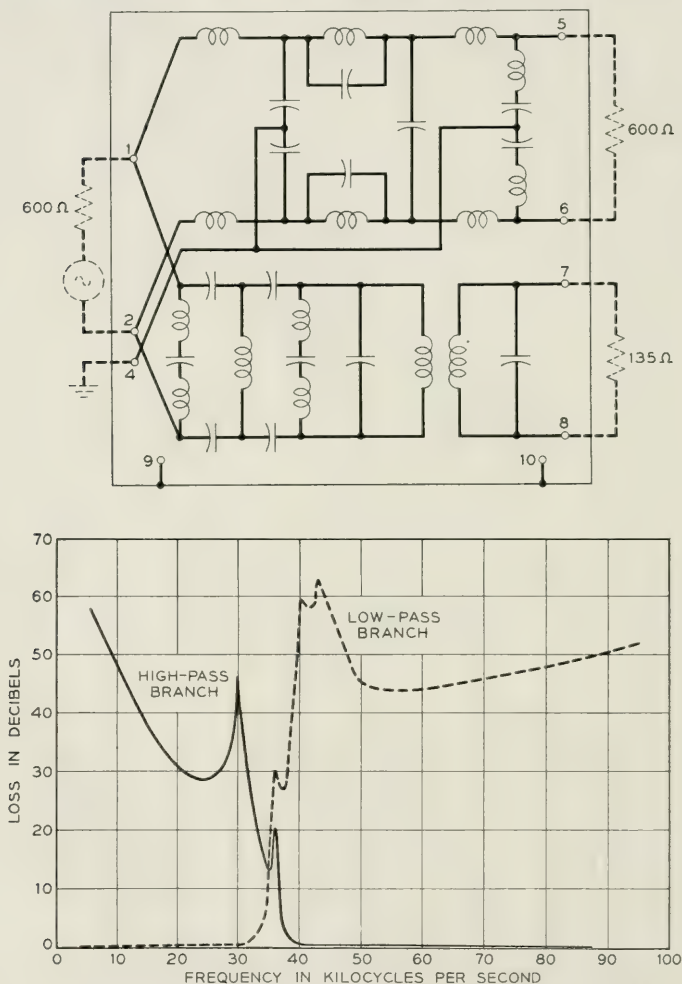


Fig. 22—Typical line filter characteristics (G1, 537A or 538A).

filters. The filter construction is shown on Fig. 32. This filter employs ferrite coils and condensers, having no crystals because of percentage band width considerations. These filters employ a form of printed wiring for interconnection of components.

The terminal framework is shown on Fig. 33. This framework employs aluminum die-castings in contrast to the fabricated framework of N. This method permits the inclusion of a slide arrangement which guides the units into place, and insures proper registration of the plugs and jacks. Some units are above the framework; others are suspended from it. The same die-casting, inverted, serves both upper and lower

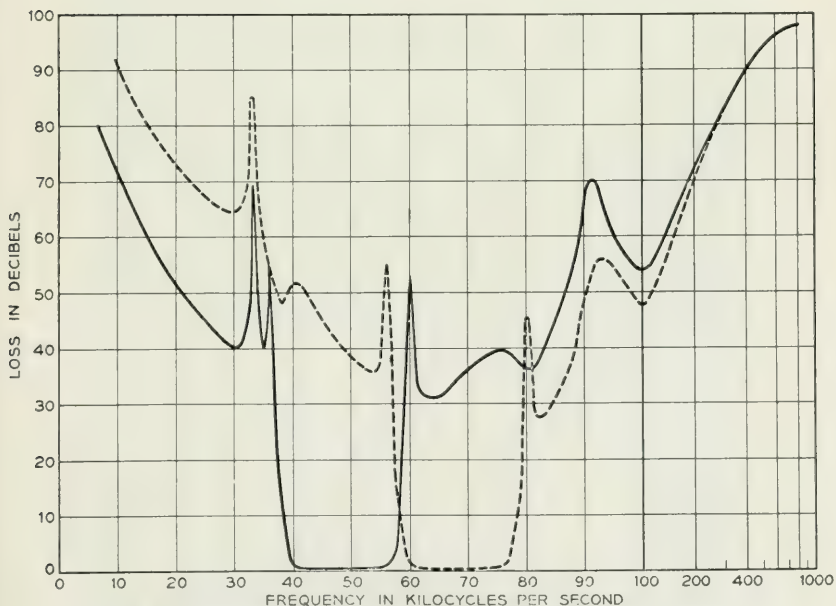
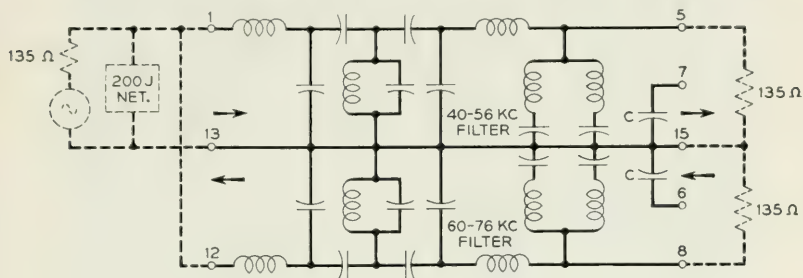


Fig. 23—Typical directional filter characteristics (C_B , 538A).

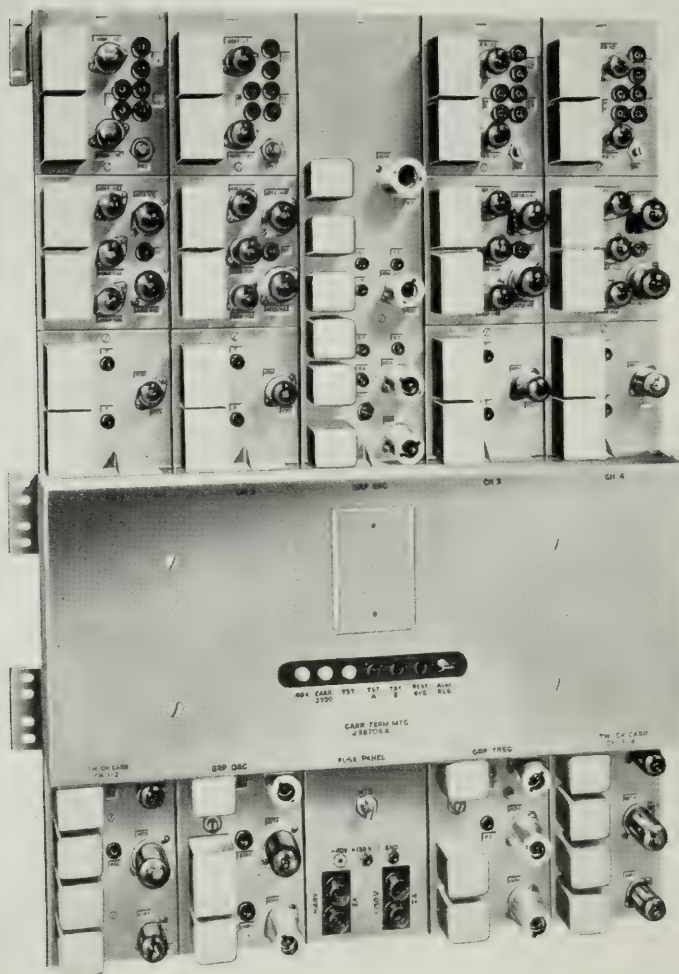


Fig. 24—Four-channel O terminal.

plug-in units. Since there is no interference between plug-in units in inserting and removing them, can covers have been eliminated. This fact, and the somewhat wider distribution of units having a high concentration of vacuum tubes, result in a relatively low temperature rise for O as compared with N. Blower facilities are not provided in the terminal.

Typical of the units suspended from the framework is the twin chan-

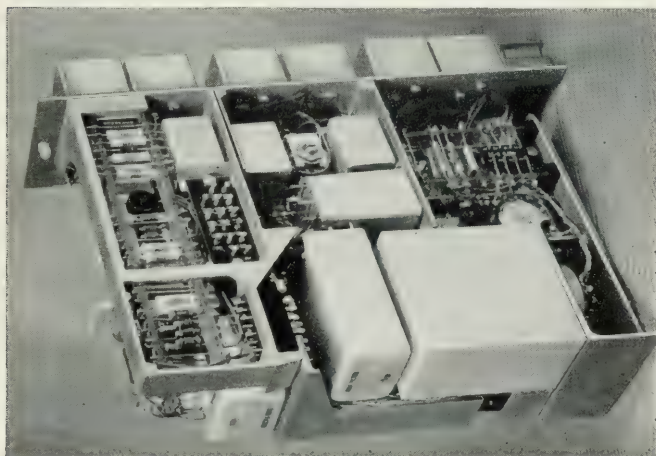


Fig. 25—Channel unit-left side view.

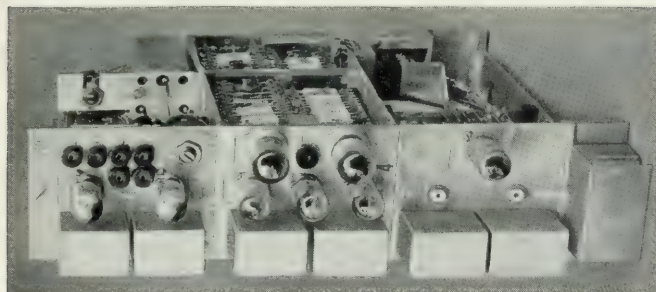


Fig. 26—Channel unit-front view.

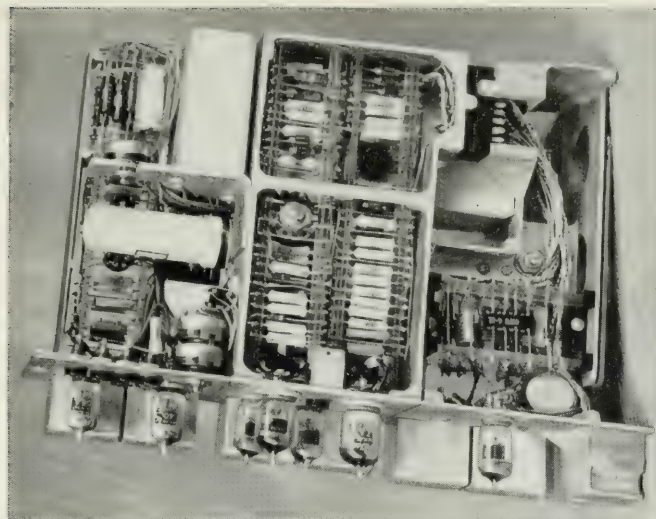


Fig. 27—Channel unit-right side view.

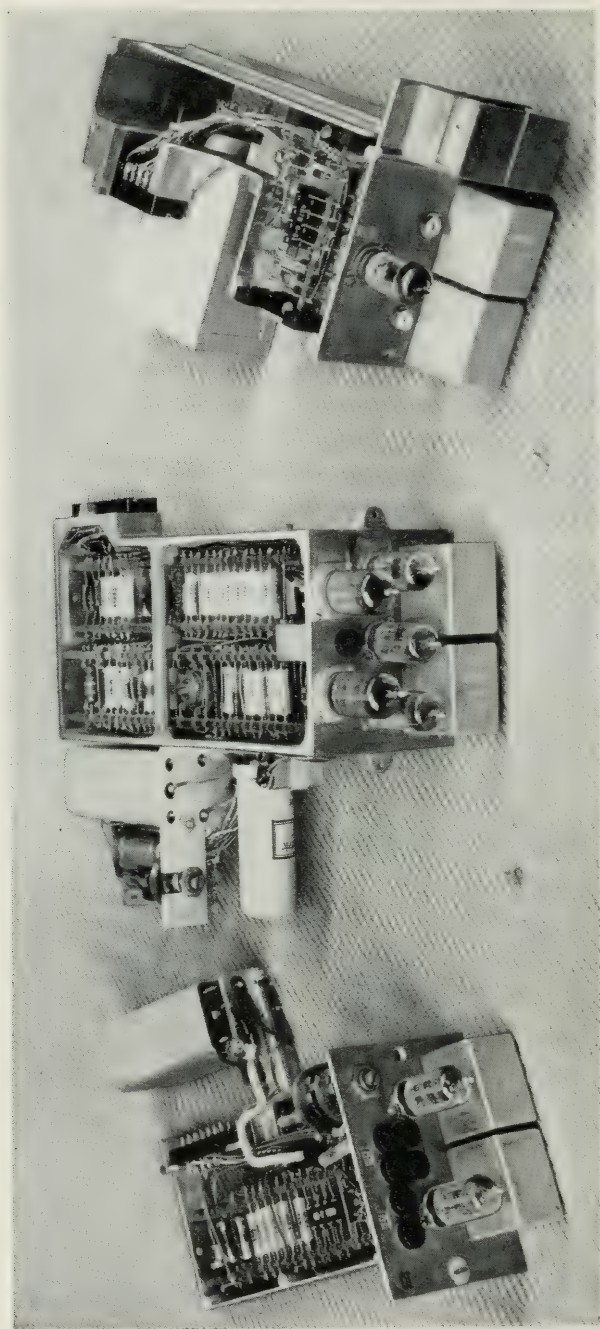


Fig. 28—Channel unit subassemblies.

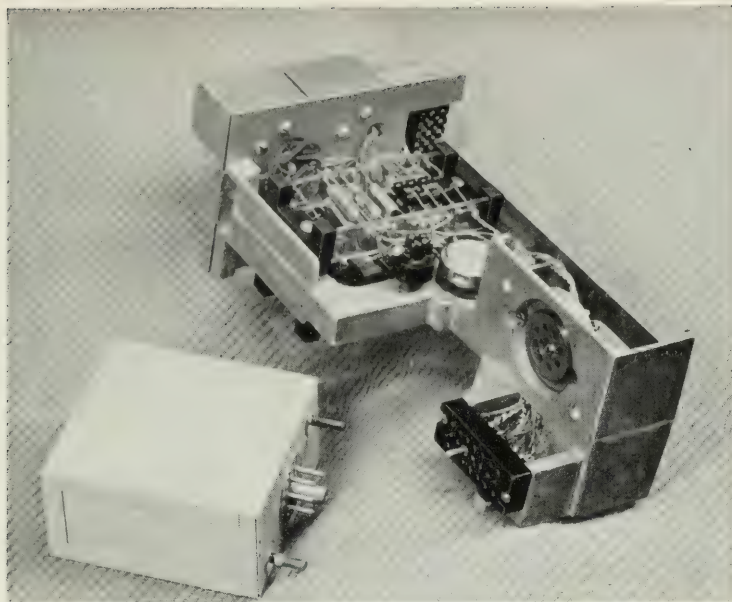


Fig. 29—Carrier subassembly and channel band filter.

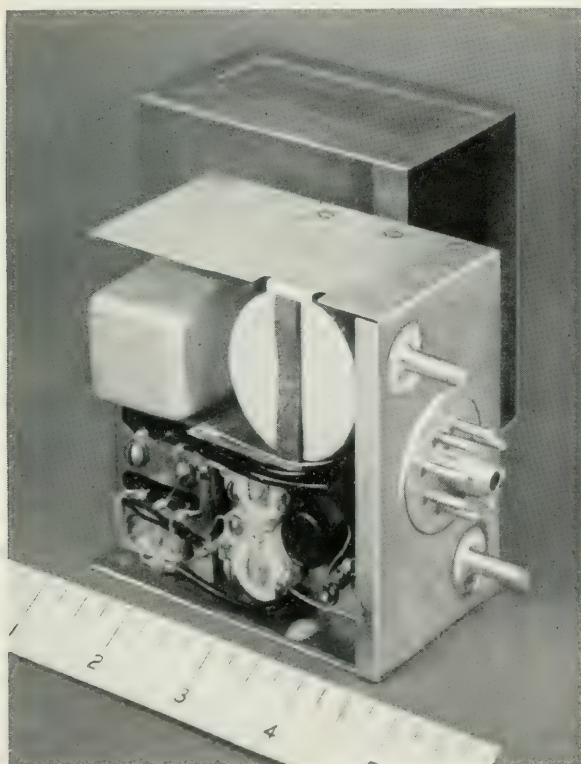


Fig. 30—Channel band filter-internal arrangement.

nel unit, shown on Fig. 34. The same basic die-casting is employed, with minor rearrangement of the die, for the two twin channel units, the group transmitting unit, and the group oscillator. The part of the slide arrangement associated with a plug-in unit is shown at the top of the twin channel unit.

All plug-in units are held in by a common cover (Fig. 24) which encloses the handles of the units. For additional support a rapid action fastener holds the tops of the channel and receiving group units.

Repeaters may be either pole mounted or placed in a central office. A group of two repeaters is shown on Fig. 35. This assembly employs a framework, shown on Fig. 36, which includes the same slide die-casting as the terminal. A central unit (also plug-in) accommodates the two

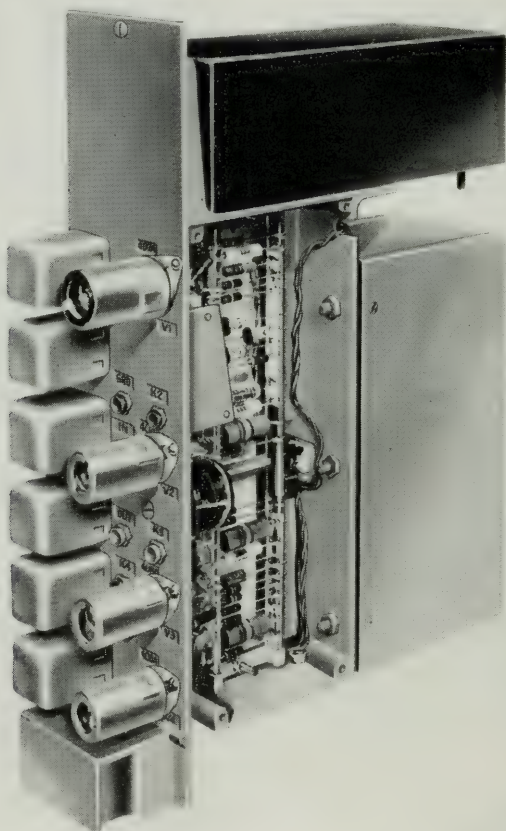


Fig. 31—Group receiving or repeater unit.



Fig. 32—Typical directional or auxiliary band filter—internal view.

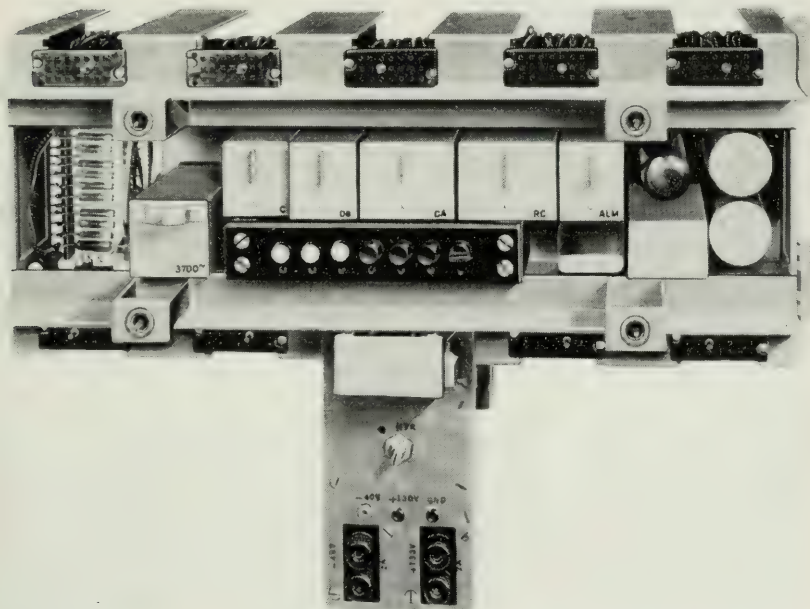


Fig. 33—Terminal framework.

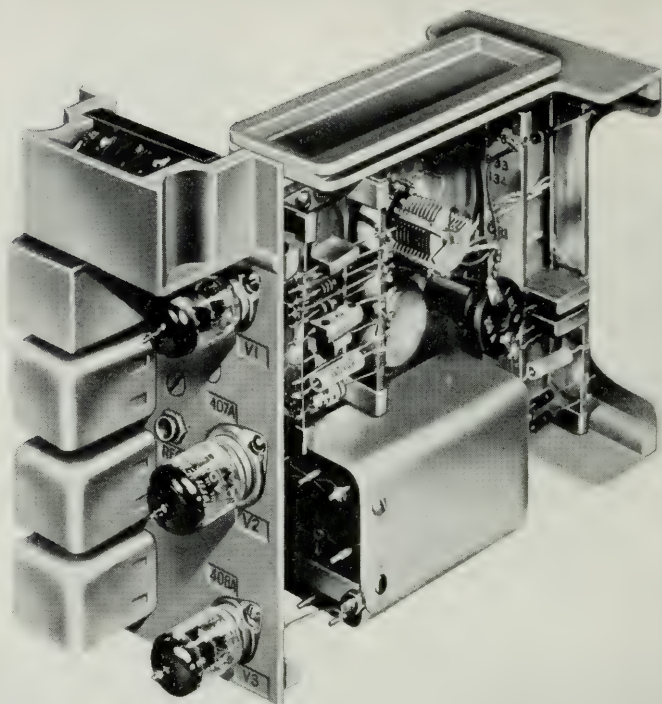


Fig. 34—Twin-channel unit.

plug-in group oscillators, which are also shown in the photograph, together with fuses, alarm lamps, etc.

Pole mounted repeaters are housed in a cabinet, similar to that used for N. Such a cabinet, equipped with four repeaters is shown on Fig. 37.

Since a maximum of four repeaters would have to be supplied by one pair of wires, it is not feasible to transmit power for the repeaters over line pairs. Instead the cabinet contains rectifiers and a line voltage regulator for obtaining 130 volts dc from commercial ac supply. For reserve power supply, a cabinet is available containing a 24-volt storage battery and a dynamotor to supply 130 volts dc to two repeaters (or two dynamotors to supply four repeaters) in case of power failure.

ALARMS

At terminals a common alarm, operating from carrier failure, performs the functions of: First, dropping all connected subscribers to prevent

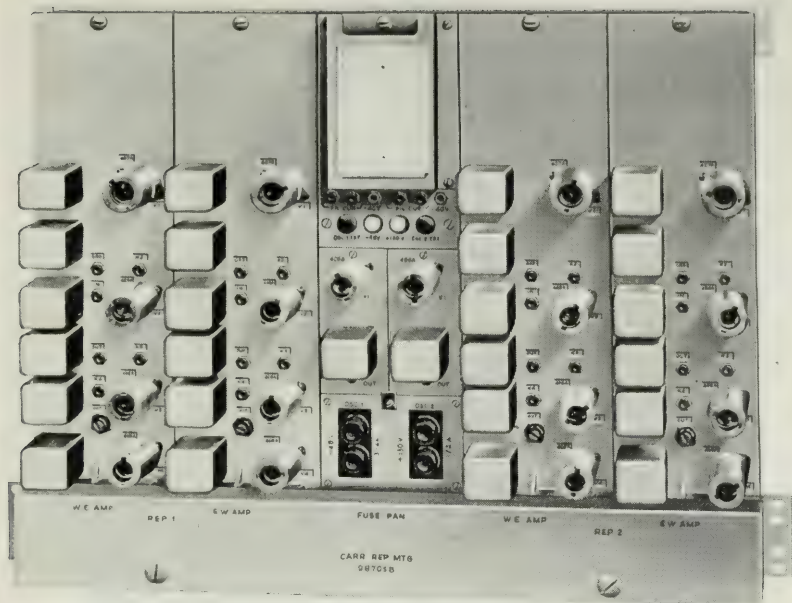


Fig. 35—Two repeater assembly.

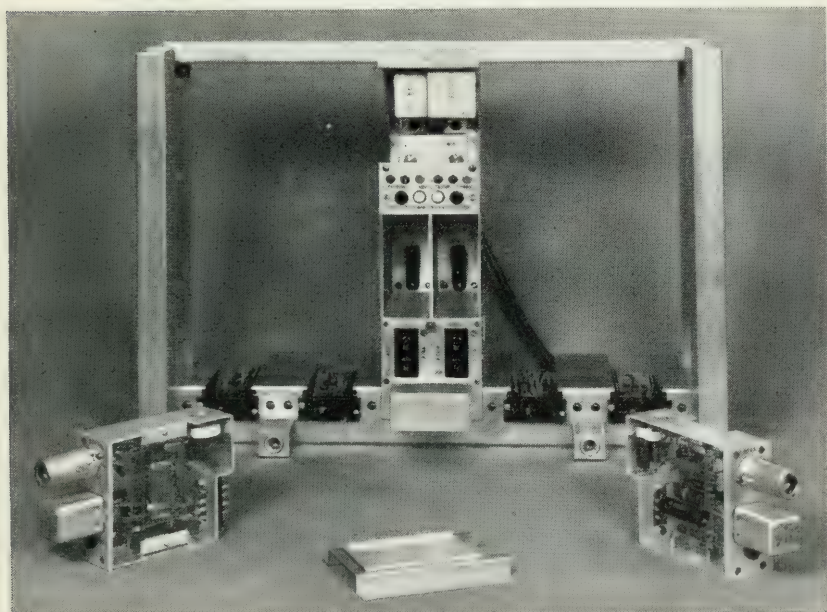


Fig. 36—Repeater framework with oscillators.

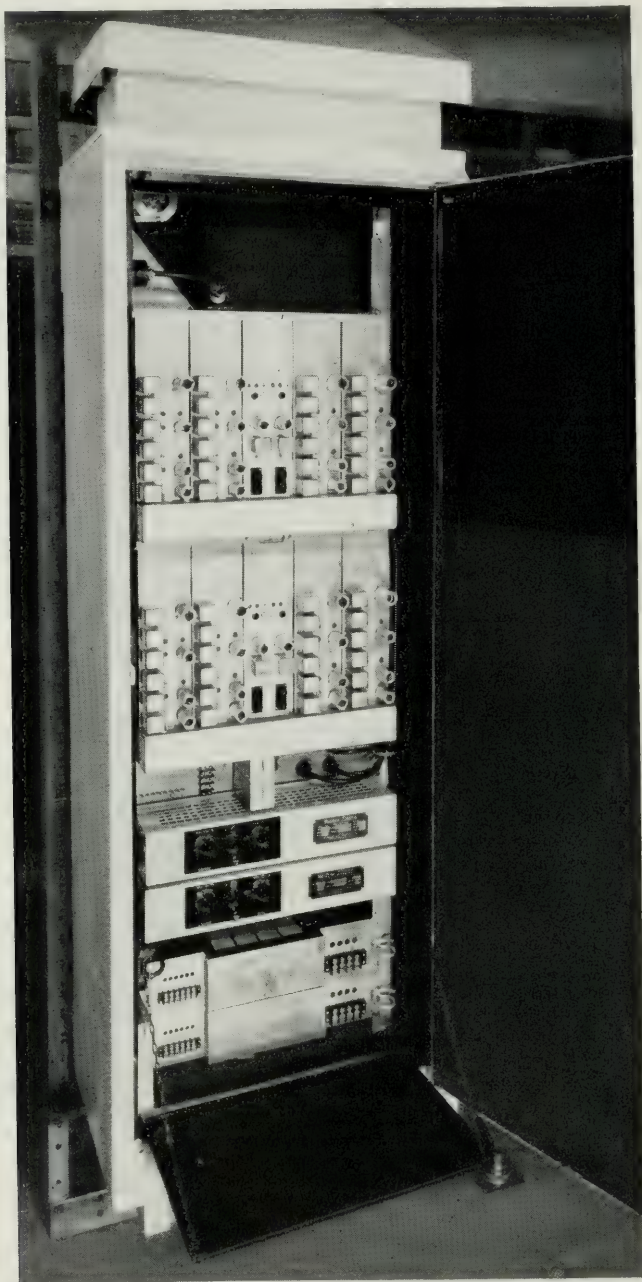


Fig. 37—Typical arrangement of pole mounted repeaters.

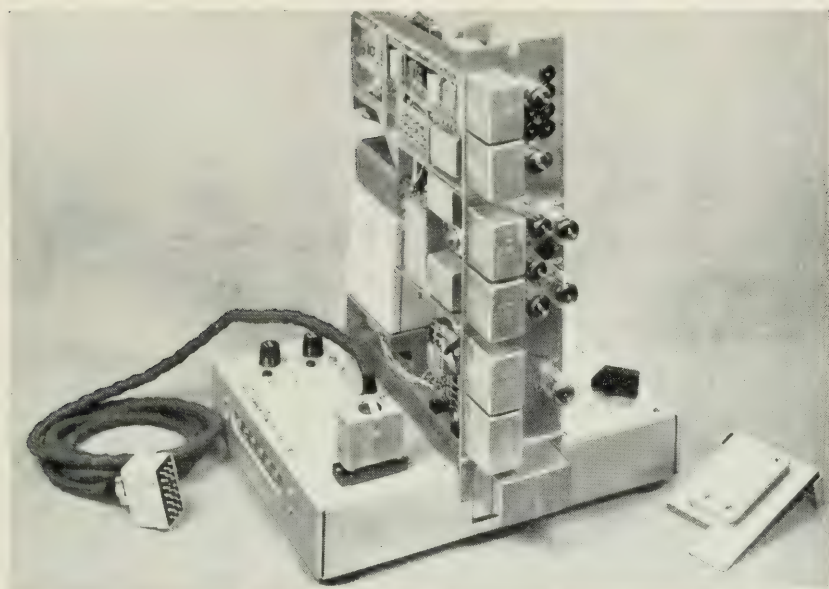


Fig. 38—Test stand.

their being held during the interval of failure; and second, to make all circuits busy at both terminals, to prevent false seizure by operators or automatic switching equipment. Since many O systems may be employed in situations where one terminal is unattended, facilities are included whereby, after failure, the system can be tested from either end, through the use of one of the signaling channels. If it is indicated that the system is operable it can be placed in service again without the necessity of a trip to the unattended terminal.

SPECIAL SIGNALING FEATURE

Arrangements are provided by which two O circuits can have their E and M signaling control leads interconnected without the use of the signaling converter, which is otherwise required. This feature is employed when two circuits are connected together on a permanent or semi-permanent basis to form a single trunk.

TESTING

To facilitate testing at terminal points a test stand (Fig. 38) has been provided which supports an O, or N, channel unit during test and adjustment. By a patch cord, the channel unit can be connected to its original framework if desired. Built in pin jacks permit bridging measurements to be made at selected points in the transmission circuit.

Efficient Coding

By B. M. OLIVER

(Manuscript received May 14, 1952)

This paper reviews briefly a few of the simpler aspects of communication theory, especially those parts which relate to the information rate of and channel capacity required for sampled, quantized messages. Two methods are then discussed, whereby such messages can be converted to a "reduced" form in which the successive samples are more nearly independent and for which the simple amplitude distribution is more peaked than in the original message. This reduced signal can then be encoded into binary digits with good efficiency using a Shannon-Fano code on a symbol-by-symbol (or pair-by-pair) basis. The usual inefficiency which results from ignoring the correlation between message segments is lessened because this correlation is less in the reduced message.

INTRODUCTION

The term coding, as applied to electrical communication, has several meanings. It means the representation of letters as sequences of dots and dashes. It means the representation of signal sample amplitudes as groups of pulses having two or more possible amplitudes as in pulse code modulation. Lately, it has also come to be the generic term for any process by which a message or message wave is converted into a signal suitable for a given channel. In this usage single-sideband modulation, frequency modulation and pulse code modulation are examples of encoding procedures, while microphones, teletypewriters and television cameras are examples of encoding devices.

This is a nice concept, but it is useful to distinguish between two classes of encoding processes and devices: those which make no use of the statistical properties of the signal, and those which do. In the first class, the encoding operation consists simply of a one-to-one conversion of the message into a new physical variable, as a microphone converts sound pressure into a proportional voltage or current, or of the one-to-one remapping of the message into a new representation without regard to probabilities, as by ordinary amplitude, frequency or pulse code modula-

tion. In ordinary PCM for example, the message samples are converted into groups of on-or-off pulses. The particular combination of pulses in any group depends only upon the amplitude of the particular sample, not upon any other property of the message, and the same time is allotted to each group, regardless of the probability of that group or of the amplitude it represents. Almost all the processes and devices used in present day communication belong to this first class. In the second class, the probabilities of the message are taken into account so that short representations are used for likely messages or likely subsequences, longer representations for less likely ones. Morse code, for example, uses short code groups for the common letters, longer code groups for the rare ones.

Processes of the first class we may call non-statistical coding processes, or simply modulation or remapping processes. The time of transmission is the same for all messages of the same length, and all messages are handled by the system with equal facility (or difficulty). These processes require no memory and have a small and constant delay. They are inefficient in their use of channel capacity.

Processes of the second class we may call statistical encoding processes. These processes in general require memory. The time of transmission of messages of the same length may be different so that if messages are to be accepted and delivered by the system at constant rates, variable delays may be necessary at the sending and receiving ends. They are more efficient in their use of channel capacity. It is with this second type of process that this paper is concerned, although processes of the first type may be used as component steps. Thus we consider systems of the type shown in Fig. 1, with the accent on the word "efficient".

TRANSMISSION CIRCUITS AND THEIR VOCABULARIES

Communication circuits or channels can, of course, differ in many respects. Either the peak signal power or the average signal power may be limited. The transmission may be uniform over the band or vary with frequency; it may be constant or subject to selective fading. The noise may be gaussian thermal or shot noise uniform across the band, or peaked at some frequency; or it may be largely impulse noise or erratic



SIGNAL = EFFICIENT DESCRIPTION OF MESSAGE

Fig. 1—Reversible statistical encoding.

static discharges. The best type of signal for one channel may be very poor for another.

In the following sections it is assumed that the channel transmission characteristic is flat in amplitude and delay over a definite band and zero outside. It is also assumed that the channel has a definite peak signal power limitation, and that the noise is white gaussian noise. Such a channel is no mere academic ideal. It is in fact quite closely approached in practice by many circuits. Moreover, the conclusions based on these assumptions can usually be modified or extended to other actual cases, such as that of noise with non-uniform spectral distribution (as for example the coaxial cable).

If the bandwidth of the channel is W , we can (using single sideband modulation, if necessary) transmit over it without distortion from frequency limitation signals containing frequencies from 0 to W (or $-W$ to W in the Fourier sense). Such a wave can assume no more than $2W$ independent amplitudes per second. Any set of samples of the wave taken at regular intervals $\frac{1}{2W}$ serves to specify the wave completely.

The wave may be thought of as a series of $(\sin x)/x$ pulses centered on the samples and of proportional height, and indeed the wave may be reconstructed from the samples in this fashion. This is the well-known sampling theorem¹. Thus a message source of bandwidth W can supply at most $2W$ independent symbols (samples) per second, and this same number can be transmitted as overlapping, but independently distinguishable pulses by a circuit of bandwidth W .

Since, as will appear later, channels which are to transmit signals resulting from efficient statistical encoding must be relatively invulnerable to noise, we shall assume that the pulses on the channel are quantized. This allows regenerative repeater to be used to eliminate the accumulation of noise¹. If there are b quantizing levels, and if the levels are sufficiently separated so that the probability of noise causing incorrect readings is negligibly small, then the capacity of the channel in bits/sec is²

$$C = 2W \log_2 b. \quad (1)$$

Such a circuit talks in an alphabet of b "letters" and uses a language in which all combinations of these letters are allowed. There are no forbidden or impossible "words". The circuit has a vocabulary of b one-letter words, b^2 two-letter words, b^n n -letter words. The basic inefficiency in present day electrical communication is that we build circuits with unrestricted vocabularies and then send signals over them which

use only a tiny fraction of this vocabulary. If all the letters of the written alphabet were used with equal probability and if all combinations of letters were allowed, then many words which are now long could be made shorter, and written text would be less than one third as long as English. Similarly, if we could arrange to let our circuits use their entire vocabulary with equal probability, they could describe our messages with much less time (or bandwidth) on the average.

EXCHANGE OF BANDWIDTH AND SIGNAL TO NOISE RATIO

It was the advent of wide band FM, and other modulation methods which exchange bandwidth for signal-to-noise ratio, which revealed the inadequacy of earlier concepts of information transmission and ultimately led to the development of modern communication theory, or information theory².

One of the more familiar results of this theory is the expression for the maximum capacity of a channel disturbed by white noise:

$$C = W \log_2 \left(1 + \frac{P}{N} \right) \quad (2)$$

in which C is the capacity in bits/sec, W is the bandwidth and P/N the ratio of average signal power to average noise power. This capacity can only be approached, never exceeded, and is only reached when the signal itself has the statistics of a white noise. The expression sets a limit for practical endeavor, and also gives the theoretical rate of exchange between W and P/N .

A practical quantized channel, operated so that the loss of information due to incorrectly received levels is negligible requires about 20 db more peak signal power than the average signal power of the ideal channel to attain the same capacity¹. However, bandwidth and signal-to-noise ratio are still exchanged on the same basis. For example, a satisfactory television picture could be sent over a channel with, say, 100 levels. This would require a (peak) signal to rms noise ratio of some $40 + 20 = 60$ db. The bandwidth could be halved by a sort of reverse PCM: by using one pulse to represent two picture elements. But there are 10,000 combinations of two samples each of which can have any of 100 values. Hence the new combination pulse would need 10,000 distinguishable levels and this would require a signal to noise ratio of $80 + 20 = (2 \times 40) + 20 = 100$ db.

It is evident that while bandwidth compression by non-statistical or straight signal remapping means is not an impossibility, it is neverthe-

less impractical when the signal to noise ratios are already high. What we should really try to do is make our descriptions of our messages more efficient so that less channel capacity is required in the first place. The saving can then be taken either in bandwidth or in signal-to-noise ratio, whichever fits the requirements of our channels best.

MESSAGES

Messages can either be continuous waves like speech, music, or television; or they can consist of a succession of discrete characters each with a finite set of possible values, such as English text. Because a finite bandwidth and a small added noise are both permissible, continuous signals can be converted to discrete signals by the processes of sampling and quantizing¹. This permits us to talk about them as equivalent from the communication engineering viewpoint. Since many of the principles which follow are easier to think of with discrete messages and since quantization of the channel is assumed for reasons already stated, we shall think of our messages as always being available in discrete form.

Let S = the symbol (or sample) rate of the message

$W_0 = \frac{S}{2}$ = the original bandwidth of the continuous message

ℓ = number of quantizing levels.

Then if all the message samples were independent and if all quantizing levels were equally likely, the information per sample would be

$$H_0 = \log_2 \ell \text{ bits} \quad (3)$$

the information rate would be

$$H'_0 = S \log_2 \ell \text{ bits/sec} \quad (4)$$

and the message would use the full capacity of a channel with ℓ quantizing levels, and bandwidth $S/2$. Or by remapping k message samples (with the ℓ possible levels) into $\left(\frac{\log \ell}{\log b}\right)k$ samples, a channel with b levels and bandwidth $W = S/2 \left(\frac{\log \ell}{\log b}\right)$ could be loaded to full capacity.

However, it is not true that the successive samples of typical messages are independent, nor is it true that the various sample amplitudes are in general equiprobable. If these things *were* true, speech and music would sound like white noise, pictures would look like the snowstorm

a TV set produces on an idle channel. Written text would look like WPEIPTNKH WFIOZ—: a random sequence of letters. The statistics of the message, in particular the correlations between the various samples, greatly reduce the number of sequences of given length which are at all likely. As a result the information rate is less, and fewer bits per second are required to describe the average message.

A sequence of M binary digits can describe any of 2^M possible messages. Conversely any of N messages can be described by $\log_2 N$ binary digits. The information rate, H , of a message source is therefore given by

$$H = \lim_{n \rightarrow \infty} \frac{\log N}{n} \text{ bits/symbol}$$

where N = number of message sequences of length n . If the successive symbols of the message are *independent* but *not equiprobable*, then a long sequence will contain x_1 symbols of type 1, x_2 of type 2, etc. The number of possible combinations of these symbols will be

$$N = \frac{n!}{\prod_j x_j!},$$

$$\text{so that } \log N = \log n! - \sum_j \log x_j!$$

For large enough n , all the x_j will be large also and we may write, by Stirling's approximation

$$\log N \rightarrow \log \sqrt{2\pi n} + n \log n - n - \sum_j [\log \sqrt{2\pi x_j} + x_j \log x_j - x_j]$$

But since $\sum x_j = n$, and since for large n , $x_j \rightarrow p(j)n$ where $p(j)$ is the probability of the j^{th} symbol, we have

$$\log N \rightarrow \log \sqrt{2\pi n} + n \log n - n \sum_j \log \sqrt{2\pi x_j} - n \sum_j p(j) \log p(j) - n \log n + n$$

$$H_1 = \lim_{n \rightarrow \infty} \frac{\log N}{n} = - \sum_j p(j) \log p(j) \quad (5)$$

which is the expression Shannon derives more rigorously². H_1 is a maximum when all the $p(j)$ are equal to $1/\ell$. Then $H_1 = \log_2 \ell = H_0$. The more unequal the $p(j)$, i.e., the more peaked the probability distribution, the smaller H_1 becomes.

If the successive samples are not independent, the message source will pass through a sequence of states which are determined by the past of the message*. In each state there will be a set of *conditional* probabilities describing the choice of the next symbol. If the state is i and the conditional probability (in this state) of the next symbol being the j^{th} is $p_i(j)$, then the information produced by this selection is

$$H_i = -\sum_j p_i(j) \log p_i(j). \quad (6)$$

The average rate of the source is then found by averaging (6) over all states with the proper weighting; thus

$$H = \sum_i p(i) H_i = -\sum_i p(i) \sum_j p_i(j) \log p_i(j). \quad (7)$$

The greater the correlation between successive symbols or samples of a message, the more peaked the distributions $p_i(j)$ become on the average, and this results in a lower value for H . As Shannon points out, the information rate of a source, as given by (7), is simply the average uncertainty as to the next symbol when all the past is known. But in a properly operating communication channel the past of the message is available at both ends, so that it should be possible to signal over the channel at the rate H bits/message symbol, rather than H_0 as we now do. In present day communication systems we ignore the past and pretend each sample is a complete surprise.

By completely efficient statistical coding it should be possible to reduce the required channel capacity by the factor H/H_0 . Whether or not this improvement can be actually reached in practice depends upon the amount of past required to uniquely specify the state of the message source. If long range statistical influences exist, then long segments of the past must be remembered. If there are m symbols in the past which determine the present state and each symbol has ℓ possible values, there will be ℓ^m states *possible* (although only 2^{mH} of these are at all probable for large m). If m is large the number of possible states becomes fan-

* In a philosophical sense the state of a message source may be dependent on many other factors besides the past of the message. If the source is a human being, for example, the state will depend on a large number of intangibles. If these could really be taken into account the resulting H for the message might be quite low. If the universe is strictly deterministic one might say that H is "really" always zero. When we describe the drawing of balls from the urn in terms of probabilities, we admit our ignorance as to the exact detail of the mixing operation which has occurred in the urn. Likewise the information rate of a source is a measure of our ignorance of the exact state of the source. From a communication engineering standpoint, the knowledge of the state of the source is confined to that given by the past of the message.

tastically large and complete statistical encoding becomes an economic impossibility if not a technical one.

Let B_i^k be a particular combination (the i^{th}) of k symbols in the past of the message. Each of these combinations at least partially determines the state of the system. Hence we can write an approximation to (7):

$$F_k = - \sum_i p(B_i^k) \sum_j p_{B_i^k}(j) \log p_{B_i^k}(j) \quad (8)$$

$F_k \rightarrow H$, as $k \rightarrow \infty$. If only m symbols in the past influence the present state, then k need only be as great as m , in order that $F_m = H$. In any case the sequence F_1, F_2, \dots, F_k is monotone decreasing. Naturally one should always pick the k symbols in the past which exert the greatest effect upon the present state, i.e. which cause $p_{B_i^k}(j)$ to be as highly peaked as possible, on the average. In English these would be the immediately previous letters; in television, the picture elements in the immediate space-time vicinity of the present element.

Suppose we break the message up into blocks of length k . Each of these blocks may be considered to be a character in a new (and huge) alphabet. If we ignore any influences from previous blocks, i.e. if we consider the blocks to be independent, then the information per block will be simply

$$- \sum_i p(B_i^k) \log p(B_i^k). \quad (9)$$

Since there are k symbols per block, the information per symbol, G_k is

$$G_k = - \frac{1}{k} \sum_i p(B_i^k) \log p(B_i^k). \quad (10)$$

As $k \rightarrow \infty$, $G_k \rightarrow H$, since the amount of statistical influence ignored (between blocks) becomes negligible compared with that taken into account.

If d is the number of binary digits required to specify a message n symbols long, then as $n \rightarrow \infty$, $d/nH \rightarrow 1$. For large n there are thus 2^{nH} messages which are at all likely out of $2^{nH_0} = \ell^n$ possible sequences (in an ℓ letter alphabet). The probability that a purely random source will produce a message (i.e., a sequence with all the proper statistics) is therefore

$$p \cong 2^{-n(H_0 - H)} \quad (11)$$

for large n . Even if $H_0 - H$ is small, $p \rightarrow 0$ rapidly for large n . This is why white noise never produces anything resembling a picture on a television screen, for instance. For in television signals, $H_0 - H > 1$

even for very complicated picture material, and $n \cong 250,000$ for a single frame.

As given by (11), p , also represents the fraction of the possible signals on a channel of ℓ levels which are likely ever to be used by messages of length n without statistical encoding.

STATISTICALLY MATCHED CODES

Since a sequence of binary digits can be remapped by a non-statistical process into a channel with b quantizing levels, or indeed into a wide variety of other signalling alphabets, it suffices to consider statistical coding processes and codes which reduce the message to a sequence of binary digits. An efficient code is then one for which the average number of binary digits, H_c , per message symbol lies between H_0 and H . As the efficiency increases $H/H_c \rightarrow 1$, so this ratio may be taken as an efficiency index. With highly efficient processes, the sequences of binary digits produced will have little residual correlation, i.e., they will be nearly random sequences. Since the encoding process must be reversible the receiver must be able to recognize the beginnings and ends of code groups. Since we have at our disposal only zeros and ones, the divisions between code groups must either be marked by a special code group reserved for this purpose, or else the code must have the property that no short code group is duplicated as the beginning of a longer group.

A code which satisfies this latter requirement and which is capable of unity efficiency is the so-called Shannon-Fano code, developed independently by C. E. Shannon of Bell Telephone Laboratories and R. M. Fano of the Massachusetts Institute of Technology. This code is constructed as follows: One writes down all the possible message sequences of length k in order of decreasing probability. This list is then divided into two groups of as nearly equal probability as possible. One then writes *zero* as the first digit of the code for all messages in the top half, *one* as the first digit for all messages in the bottom half. Each of these groups is again divided into two subsets of nearly equal probability and a zero is written as the second digit if the message is in the top subsets, a one if it is in the bottom. The process is continued until there is only one message in each subset. Fig. 2a shows the code which results when this process is applied to a particularly simple probability distribution $p(B_i^k) = (1/2)^i$. Here each code group is a series of ones followed by a zero. The receiver knows a code group is finished as soon as a zero appears. Although the longer groups contain mostly ones, their probability is less and on the average as many zeros are sent as ones.

MESSAGE		CODE	STEP
NO.	PROB.		
1	$\frac{1}{2}$	0	(1)
2	$\frac{1}{4}$	1 0	(2)
3	$\frac{1}{8}$	1 1 0	(3)
4	$\frac{1}{16}$	1 1 1 0	(4)
5	$\frac{1}{32}$	1 1 1 1 0	(5)
6	$\frac{1}{64}$	1 1 1 1 1 0	
7			

(a)

MESSAGE		CODE	STEP
NO.	PROB.		
1	$\frac{1}{4}$	0 0	(2)
2	$\frac{1}{4}$	0 1	(1)
3	$\frac{1}{8}$	1 0 0	(3)
4	$\frac{1}{8}$	1 0 1	(2)
5	$\frac{1}{16}$	1 1 0 0	(4)
6	$\frac{1}{16}$	1 1 0 1	(3)
7	$\frac{1}{32}$	1 1 1 0 0	

(b)

MESSAGE		CODE	STEP
NO.	PROB.		
1	$\frac{1}{8}$	0 0 0	(3)
2	$\frac{1}{8}$	0 0 1	(2)
3	$\frac{1}{8}$	0 1 0	(3)
4	$\frac{1}{8}$	0 1 1	(1)
5	$\frac{1}{8}$	1 0 0	(3)
6	$\frac{1}{8}$	1 0 1	(2)
7	$\frac{1}{8}$	1 1 0	(3)
8	$\frac{1}{8}$	1 1 1	

(c)

Fig. 2—Shannon-Fano codes for three different distributions. The successive bisections are indicated by the dashed lines and the number gives the step at which that bisection took place.

If the successive message segments are independent, the code will generate a random sequence of zeros and ones. Fig. 2b shows the code which results with another distribution. Here the termination of each code group is more complicated but the non-duplicative property exists so the receiver can still identify the groups. Fig. 2c shows the code which results when all the $p(B_i^k)$ are equal. It is the ordinary binary code.

The length of each code group is equal to $\log 1/p(B_i^k)$, for the cases shown in the figures. This is true in general so long as it is possible to divide the list into subgroups which are of exactly equal probability.

When this is not possible, some code groups may be one digit longer as Shannon shows. The average number of digits per message symbol using this code is therefore given by

$$-1/k \sum_i p(B_i^k) \log p(B_i^k) \leq H_c \leq -1/k \sum_i p(B_i^k) [-1 + \log p(B_i^k)]$$

$$G_k \leq H_c \leq G_k + 1/k.$$

For large k , $H_c \rightarrow G_k \rightarrow H$ and the efficiency approaches unity. With small k , H_c increases both because the smaller list of messages cannot be so accurately divided repeatedly into equal probability subsets (so-called "granularity" trouble), and also because more statistics are ignored between the shorter blocks.

The ordinary binary code provides a statistical match between message source and channel only if the various message blocks B_i^k have equal probability $p(B_i^k) = 1/2^n$, and are mutually independent. With $k = 1$, $p(B_i^k) = p(j)$ and the "blocks" are merely the successive symbols.

Ordinary PCM is statistically matched only to a random message source with flat distribution.

If the messages from a source are characterized by frequent long runs of symbols of the same type (e.g., long runs of zeros) an obvious saving is possible by sending the value of the symbol only once, together with a code group which gives the length of the run. This is commonly known as run length coding. The remaining sections of the message (between runs) may then either be sent directly (i.e., merely remapped by a non-statistical process) or they may be encoded by some other statistical process, if this seems warranted. In the latter case we have a mixed coding procedure. The codes representing run lengths must either be set apart from the remainder of the signal by "punctuating" codes, or identifiable by some distinguishing characteristic.

Run length coding may be generalized to take care of other common sequences besides runs of a single symbol. Any commonly occurring sequence of symbols may be considered a "run" and treated in the same fashion. More complicated code groups will be required to specify the type of run, if a large variety is accommodated this way. Ultimately, the distinction between this type of coding and Shannon-Fano coding becomes rather nebulous, especially if a fixed maximum length of run is permitted, for then all possible messages of this length may be considered "runs" and simply encoded by the Shannon-Fano code.

No optimal general solution of the coding problem is known. That is, one cannot say in all cases exactly what coding procedure one should use with a given message source to produce the most efficient encoding for a given complexity of apparatus. Several procedures have been devised which seem suitable for certain types of messages and these are discussed in the following sections.

n -GRAMMING

The application of the Shannon-Fano code to a block of k symbols of a message in an ℓ letter alphabet requires that ℓ^k different codes be used. The receiver must be able to recognize each of these and to regenerate the proper message block when a particular code is received. If ℓ is on the order of 10 to 100 as is typically the case, we very quickly run out of room to house the receiver and money to build it with. On the other hand, if k is small, say on the order of 1 to 3, considerable statistical information between blocks is ignored. These considerations led to the development of a class of encoders known as n -grammers. The name stems from the fact that they operate on the n -gram statistics of the

message, to produce a reduced signal having more nearly independent symbols, but (in return) a highly peaked simple probability distribution which allows savings with Shannon-Fano coding on a symbol-by-symbol ($k = 1$) basis.

The simplest member of this class is the monogrammer. It is basically merely a re-ordering device. The operation may be best understood by the following example. Suppose someone supplied us with English text encoded into a quantized pulse signal as follows:

Symbol	Pulse height
Space	0
<i>A</i>	1
<i>B</i>	2
<i>C</i>	3
<i>D</i>	4
etc.	etc.

Now the letter frequencies in English are shown in Fig. 3. Merely to save average power in our channel we might wish to convert this signal into one in which the pulse height is not alphabetical, but in which the most common symbol is sent as a pulse of zero height, the next most common as a pulse of unit height, etc. In other words, we would like the following representation:

Symbol	Pulse height
Space	0
<i>E</i>	1
<i>T</i>	2
<i>A</i>	3
etc.	etc.

The device shown in Fig. 4 will accomplish this translation. The original signal is applied to the vertical deflecting plates of a cathode ray tube. The rest position of the spot corresponds to "space", i.e. no pulse. A pulse one unit high deflects the spot to *A*, a pulse two units high deflects the spot to *B*, etc.

Now in front of these spot positions we place a number of light attenuating filters. In front of the "space" position we place an opaque mask. Hence when the spot is deflected to "space" the photocell receives no light and no pulse is sent. In front of the "*E*" position we place a mask having one unit of transmission. So although *E* is received as a pulse 5 units high, it is sent as a pulse of unit height. In front of the

"T" position we place a mask with two units transmission, and so on. The signal amplitudes as received are thus re-ordered in the desired fashion.

The resulting signal has lower average power and this can sometimes be an advantage, particularly if several such signals are to be sent over a common channel by frequency division. In this case the extreme rarity of occurrence of high peak powers on all channels simultaneously means

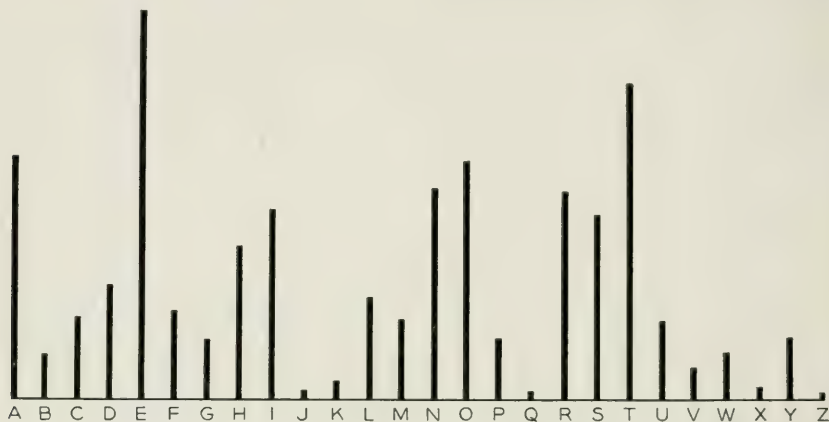


Fig. 3—Letter frequencies in English.

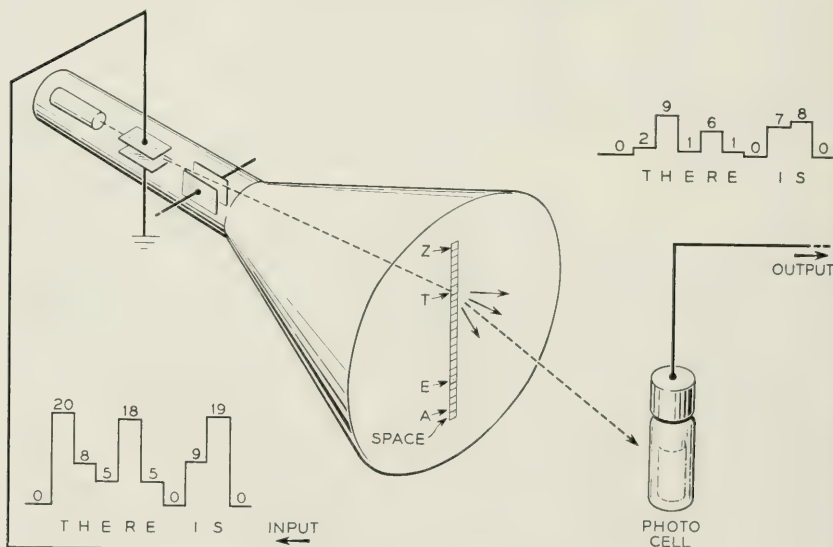


Fig. 4—The "Monogrammer."

that the system can be designed to have a lower *peak* power capacity. The signal out of the monogrammer can be remapped into binary digits using a Shannon-Fano code, pulse by pulse. However, this could have been done equally well with the original signal merely by rearranging the code groups in the coder tube. It is when we extend the principle to digrams and trigrams that the potentialities of the system become evident.

We can easily take account of the influence of the preceding message symbol. To do this we apply the signal to the vertical plates as before, and to the horizontal plates we apply the signal delayed by an amount equal to the time between successive pulses as shown in Fig. 5. Thus the beam is deflected *vertically* by the *present* message symbol, and *horizontally* by the *previous* message symbol. Whereas before we used a single column of optical filters chosen in accordance with the simple probabilities of the letters, we now have 27 columns, one for each letter and one for the space. The filters in each column are chosen in accordance with the *conditional* probabilities which apply when the corresponding letter was the previous symbol. For example, in the "Q" column (last letter Q), and the "U" row (present letter U) the mask would be opaque, since U is most common after Q. In general, the transmission of cell ij , in the i^{th} column and j^{th} row, is proportional to the rank of the entry for $p_i(j)$ when the entire distribution (conditioned on i) is ordered in a monotone decreasing sequence. The amplitude distribution of the output pulses

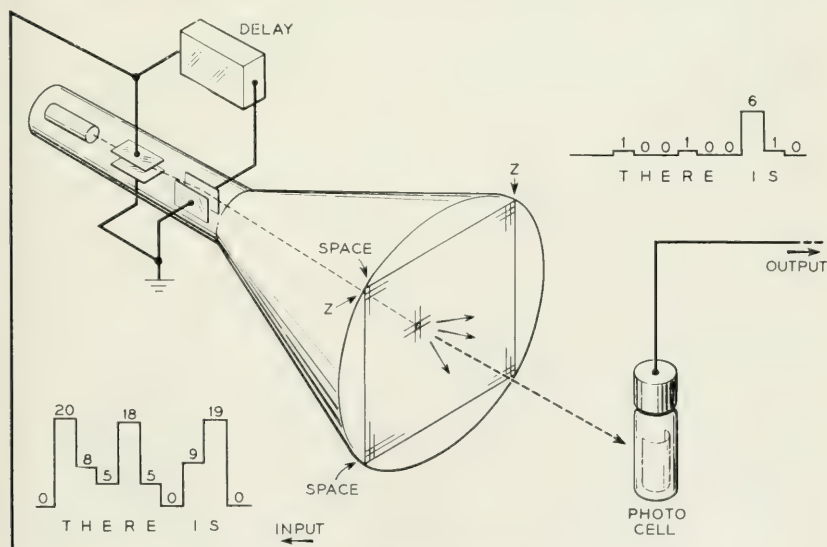


Fig. 5—The "Digrammer."

from the digrammer will be more peaked toward zero amplitude than that of the monogrammer. This is illustrated by the signals in the figures. At the receiver the same type of device, but with an inverse mask can be used to convert the signal back to its original form.

The digrammer can, with a little assistance, supply all the data required to prepare the encoding mask. If typical signals from the message source are applied to the cathode ray tube (without mask) for a long time, and a time exposure is made of the face of the tube, a lattice of spots will be obtained on the film. These spots will be dense where the high probability combinations occur and less dense elsewhere. The order of decreasing density in each column is noted, and the filter transmissions are arranged in the same order.

It is, of course, not necessary to use a phosphor, optical filters, and a photocell. An array of targets each of which connects to the appropriate tap on a load resistor might be simpler and more efficient. The cathode ray tube itself can be replaced with an appropriate diode switching network. Relay networks could be used for low-speed operation.

At the digrammer level we run out of new dimensions to use in the cathode ray tube. The principle can, however, be extended to trigramming and general n -gramming. For example, tetragramming could be accomplished by using a bank of ℓ^2 digrammers all in parallel, and all deflected by the present and previous samples. Only one of these tubes would be turned on at a time however. Which one this was would depend on the other two previous symbols of the tetragram. These (by additional delays) would be applied to the deflecting plates of a master switching tube having an array of target plates in place of a mask. Depending on the particular combination of signal samples applied to this tube, the beam would strike a particular target. The target current would then be used to turn on the beam of a particular digrammer tube, namely the one with the proper mask for that particular combination of two past symbols.

The complete array of equipment is admittedly rather staggering, but then, rather efficient coding should result. In practice it would probably be found that the masks of many of the tubes would be so similar that little gain resulted from differentiating between them. That is, the state of the message source might be nearly equivalent for several past combinations. In these cases, the group of tubes could be replaced with one having the best average mask, and the corresponding targets on the switching tube then tied together. This compromise would be particularly warranted for those tubes which were rarely used anyway. By these

tricks it should be possible to keep the growth of equipment down to something approaching 2^{nH} rather than ℓ^n .

The output signal from the n -grammer will be, as we have seen, a series of pulses with an amplitude distribution very peaked toward zero and small pulses. If ℓ , the alphabet, is large, these pulses can be efficiently encoded into a Shannon-Fano code. For small alphabets, granularity trouble can be reduced by remapping the output pulses two-by-two into pulses of base ℓ^2 , and then encoding these into the Shannon-Fano code.

The output signal from the n -grammer with English text as the input message is a pulse amplitude representation of the type of "reduced text" one gets by using running n -gram prediction on English, as described by Shannon³.

More efficient encoding would result if the properly matched Shannon-Fano code for *each particular conditional distribution* were applied to the output pulses, rather than using the same code for all of them. The efficiency of the coding operation would then be close to F_n as given by (8) (take $k = n$). This would add a great deal to the complexity and with most signals it is felt the gain would be small. If all the conditional distributions were alike after ordering, the improvement would be nil.

English text was used as the message in describing the n -gramming technique to emphasize the fact that it is a powerful general method which works even when the conditional probability distributions of a message are disorderly, multimodal affairs. It is obviously suited to other types of messages as well. Its main drawback is the complexity of apparatus required.

PREDICTIVE-SUBTRACTIVE CODING

When the conditional probability distributions of a message are unimodal (or merely strongly peaked as a rule in the vicinity of a particular sample amplitude) it is not necessary to re-order the distributions in order to obtain a reduced message for coding. The distributions may then merely be shifted along the amplitude scale until their modes are near zero (or their second moments about zero are nearly minimum). This shifting can be accomplished by computing from the preceding $(n - 1)$ gram the amplitude at which this mode or mean is located, and then subtracting this computed amplitude (or the nearest quantizing level) from the actual amplitude of the present sample. The difference in each case is a symbol whose amplitude distribution is peaked in the vicinity of zero amplitude. Fig. 6 shows a block schematic of a system using pre-

dictive-subtractive coding. In an actual system the reduced signal would ordinarily be encoded into Shannon-Fano code groups before transmission over the channel.

If s_0 is the present sample amplitude, and $s_1, s_2, s_3 \dots s_n$ are previous sample amplitudes we compute a predicted value, s_p , for the present sample which is given by

$$s_p = f(s_1, s_2, \dots s_n) \pm \delta$$

where $\delta < \frac{1}{2}$ quantizing level. If the conditional probability distribution for the present sample is $p_{s_1 \dots s_n}(s_0)$, then the difference, or output, or

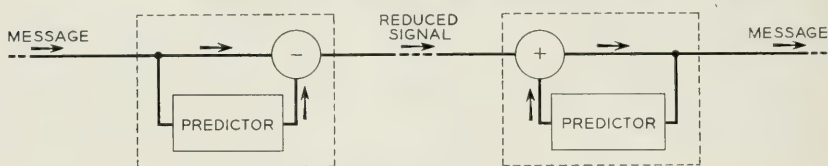


Fig. 6—Predictive-subtractive coding.

“error” signal, ϵ , will have the conditional distribution $p_{s_1 \dots s_n}(\epsilon + s_p)$ for this particular case. The simple distribution is then the weighted average over all cases, i.e.

$$p(\epsilon) = \sum p(s_1, s_2, \dots s_n) p_{s_1 \dots s_n}(\epsilon + s_p)$$

where the sum is over all combinations of $s_1, s_2 \dots s_n$.

Predictive-subtractive coding has especial merit when a simple function can be used for computing s_p . This is often the case. When the function is simply a weighted sum of the past sample amplitudes, i.e. when

$$s_p = (as_1 + bs_2 + cs_3 + \dots) \pm \delta$$

we have what is known as *linear prediction*. Of course, linear prediction can always be used, but it may not be good enough with some types of messages.

As Wiener has shown the coefficients $a, b, c \dots$ which minimize $\bar{\epsilon}^2$ are readily computed. For simplicity, assume only two message samples, s_1 and s_2 , from the past are to be used. We then have

$$\epsilon = s_0 - s_p$$

$$\epsilon = s_0 - as_1 - bs_2$$

$$\epsilon^2 = s_0^2 + a^2 s_1^2 + b^2 s_2^2 - 2as_0 s_1 + 2ab s_1 s_2 - 2b s_0 s_2$$

Now

$$\begin{aligned}\overline{s_0^2} &= \overline{s_1^2} = \overline{s_2^2} = A_0 \\ \overline{s_0 s_1} &= \overline{s_1 s_2} = A_1 \\ \overline{s_0 s_2} &= A_2\end{aligned}$$

where A_0 , A_1 , and A_2 are the values of the auto-covariance of the message wave at displacements of 0, 1, and 2 sampling periods. Thus

$$\overline{\epsilon^2} = (1 + a^2 + b^2)A_0 + 2(ab - a)A_1 - 2bA_2.$$

The autocorrelation (normalized auto-covariance) is given by $\phi_i = \frac{A_i}{A_0}$. A_0 is proportional to the average power in the message wave, so the ratio $\rho = \frac{\overline{\epsilon^2}}{A_0}$ is the ratio of the power in the error signal to the power in the original message wave. Thus:

$$\rho = (1 + a^2 + b^2) + 2(ab - a)\phi_1 - 2b\phi_2$$

$$\frac{\partial \rho}{\partial a} = 2a + 2(b - 1)\phi_1 = 0$$

$$\frac{\partial \rho}{\partial b} = 2b + 2a\phi_1 - 2\phi_2 = 0$$

from which

$$a = \frac{\phi_1(1 - \phi_2)}{1 - \phi_1^2}, \quad b = \frac{\phi_2 - \phi_1^2}{1 - \phi_1^2}.$$

With these values of a and b :

$$\rho = 1 - \phi_1^2 - \frac{(\phi_1^2 - \phi_2)^2}{1 - \phi_1^2}.$$

If $\phi_2 = \phi_1^2$, then the expressions simplify to

$$a = \phi_1, \quad b = 0, \quad \rho = 1 - \phi_1^2.$$

As can easily be shown, if $\phi(x) = e^{-\alpha|x|}$, then all the coefficients except a are zero, and a has the value $e^{-\alpha}$. In other words, if the autocorrelation function is of exponential shape, the previous sample *alone* is needed

for linear prediction. Samples before this add no further information as to the location of the mean of the conditional distributions.*

It happens that in typical television signals the autocorrelation for small displacements shows a very nearly exponential behavior. Thus linear prediction on the basis of the previous picture element alone is a natural method for television, particularly in view of the simplicity of apparatus required.

Linear prediction is easily instrumented. Fig. 7 shows in block schematic form the essentials of a linear predictor. Samples of the message are applied to a delay line. Taps along this line separated by the inter-symbol time of the message, or multiples thereof, make the desired past symbols available. The signals from these taps are merely attenuated by amounts corresponding to the coefficients $a, b, c \dots$ and added. A differential summing amplifier is shown to allow for negative coefficients, and also to accomplish the subtraction of the predicted sample amplitude from the present sample amplitude.

A complete linear predictor-subtractor is nothing but a transversal (time domain) filter whose impulse response is

$$f(t) = \delta(t) - a\delta(t - \tau) - b\delta(t - 2\tau) \dots$$

and whose equivalent frequency response is therefore

$$F(\omega) = 1 - ae^{-i\omega\tau} - be^{-2i\omega\tau} \dots$$

where τ is the delay between taps. If, for example, simple previous value prediction is used ($a = 1; b, c \dots = 0$)

$$F(\omega) = 1 - e^{-i\omega\tau} = 2i \sin \frac{\omega\tau}{2} e^{-\frac{i\omega\tau}{2}}.$$

* From the preceding expression for ρ , we see that $\rho = 0$ (i.e., perfect prediction is possible) if:

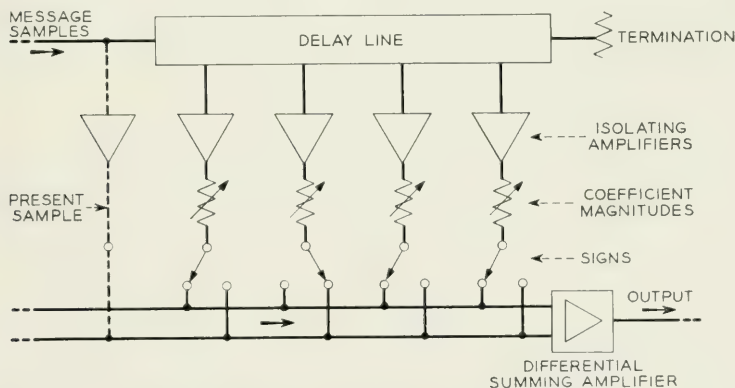
$$\begin{aligned} (\phi_1^2 - \phi_2)^2 &= (1 - \phi_1^2)^2 \\ \phi_1^2 - \phi_2 &= \pm(1 - \phi_1^2) \\ \begin{cases} \phi_2 = 1 \\ \phi_2 = 2\phi_1^2 - 1. \end{cases} \end{aligned}$$

If $\phi_2 = 1$, the message samples alternate between two independent but constant values. For this case $a = 0, b = 1$. If $\phi_2 = 2\phi_1^2 - 1$ the autocorrelation is a cosine wave so the message consists of samples of a sinusoid. In this case $a = 2\phi_1, b = -1$. If ϕ_1 is nearly unity, the sinusoid is of low frequency, and the prediction approaches "slope" prediction (i.e. extrapolation of a straight line through the last two samples).

In any case where perfect prediction is possible the wave is periodic and therefore $H = 0$.

It is often argued that linear prediction is therefore nothing more than pre-distortion (frequency-wise). If the message is unquantized and un-sampled, and if the signal from the predictor is applied to the channel as straight amplitude or single side-band modulation, the allegation is certainly true. Pre-distortion is a perfectly valid way of improving the statistical match between message, and channel, and destination as the optimum filter theory of Weiner and Lee shows. On the other hand, when the message is sampled and quantized, and when the output of the linear predictor is further encoded into a sequence of binary digits, and these are possibly remapped onto a higher base for the channel, then the information is being handled digitally throughout, and the usual reasons for a certain type of predistortion no longer apply. The best linear predictor will usually be quite different for the two cases. Even though analogue operations (such as subtraction of amplitudes) are used for convenience, the quantization makes the operation discrete and hence equivalent to a digital process.

At the beginning of this section, we were a little vague as to whether the prediction should shift the modes or the means of the conditional distributions to zero amplitude. If the object of the prediction-subtraction operation is to minimize the *power* in the error signal, then certainly the means should be shifted to zero. The coefficients as determined from the autocorrelation function do this aside from quantizing granularity. They specify an optimum least-square predictor, i.e., one which tends to minimize $\overline{\epsilon_j^2} = \sum_j j^2 p(j)$.



IF PRESENT SAMPLE IS SUBTRACTED, OUTPUT WILL
BE "ERROR" SIGNAL
IF PRESENT SAMPLE IS OMITTED, OUTPUT WILL
BE "PREDICTION"

Fig. 7—A linear predictor.

Power reduction is an index of merit when many reduced signals are to be sent by frequency division over one channel, as we have said. When the object is to reduce the channel capacity required for a single message source, then it is the upper bound entropy of the reduced signal which should be minimized, not the power. That is we want $-\sum_j p(j) \log p(j)$ to be minimized. For certain types of signals this requires the modes to be shifted to zero, although this is by no means a general rule. Shifting the modes to zero may actually increase the entropy of the "reduced" signal over that of the original message, by adding too many new symbol levels, as the example in the last section shows.

If the original message has ℓ quantizing levels, the reduced message after predictive-subtractive coding will in general contain more than ℓ levels since an error of more than $\frac{\ell}{2}$ can be made in either direction. An n -gramming operation, on the other hand, never increases the alphabet.

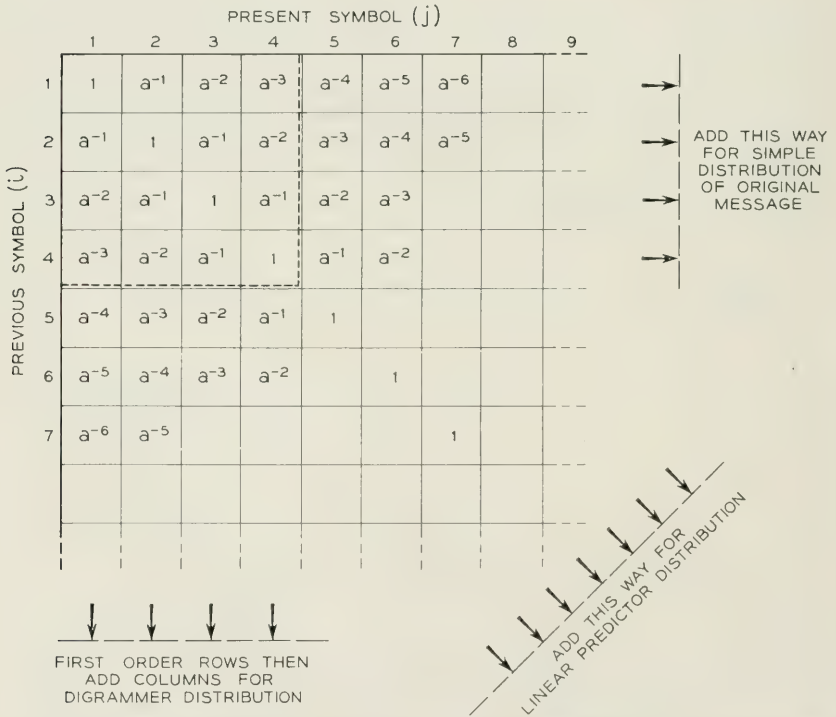


Fig. 8—Joint probability distribution (divide all coefficients by the sum over each array).

Other operations besides simple subtraction of the predicted symbol from the present symbol are of course possible. However, in most cases it would seem that if a more complicated operation were indicated, n -gramming would have provided a better start.

ILLUSTRATIVE EXAMPLE

Let us compare the operation of n -gramming and prediction-subtraction techniques on a hypothetical message. We will assume the message has digram statistics, but that longer range statistical influences either do not exist or are ignored. The statistics are then specified entirely by the joint probability distribution $p(i, j)$ of a pair of symbols. Let us assume that there are ℓ quantizing levels, and that

$$p(i, j) = K a^{-|i-j|}$$

where a is a constant > 1 , and K is given by

$$K = \left[\sum_{i,j} a^{-|i-j|} \right]^{-1}$$

K is the factor which assures that $\sum_{i,j} p(i, j) = 1$.

Thus the most likely level is that of the previous sample. A sample differing by one level is $1/a$ times as likely, one differing by m levels is a^{-m} times as likely. Figure 8 shows a plot of the *relative* values of $p(i, j)$ (neglecting the factor K). For $\ell = 4$, the total array would be the 4×4 portion enclosed by the dashed line. This sort of distribution is rather similar to those of typical television signals, as shown by preliminary measurements, although typical values of a have yet to be determined. With no statistical coding, the required channel capacity is

$$H_0 = \log_2 \ell \text{ bits/sample.}$$

If the simple distribution of individual samples is taken into account, the required channel capacity is reduced to

$$H_1 = -\sum p(i) \log p(i)$$

where

$$p(i) = \sum_j p(i, j) = \sum_i p(i, j) = p(j)$$

H_1 may be computed from the array of *relative* coefficients by adding the rows to form the sums

$$S_i = \frac{1}{K} \sum_j p(i, j) = \frac{p(i)}{K}.$$

In terms of these sums, we have

$$H_1 = \log \frac{1}{K} - K \sum_i S_i \log S_i.$$

Since, with the assumed distribution, the S_i are all nearly equal very little reduction in channel capacity is achieved by this step.

With linear prediction, the modes of the distributions ($i = j$) could be centered at zero merely by sending the difference between the present and previous sample (previous value prediction). This would give a reduced signal whose distribution may be found by adding the array along the diagonals. The required channel capacity is then given by:

$$H_L = -K \ell \log K \ell - 2 \sum_{k=1}^{\ell-1} \frac{k(\ell - k)}{a^k} \log \frac{k(\ell - k)}{a^k}$$

The distribution of the signal from a digrammer is found by rearranging each row of the table in order of decreasing probability and then adding the resulting columns. Call these sums S_d . The digrammer output will thus require a channel capacity:

$$\begin{aligned} H_0 &= - \sum_{d=1}^{\ell} K S_d \log K S_d \\ &= \log \frac{1}{K} - K \sum_{d=1}^{\ell} S_d \log S_d \end{aligned}$$

Lastly, the true rate of the source is given by

$$\begin{aligned} H &= - \sum_i p(i) \sum_j p_i(j) \log p_i(j) \\ &= \log \frac{1}{K} - H_1 + 2K \sum_{k=1}^{\ell-1} \frac{\ell - k}{a^k} k \log a \end{aligned}$$

Values for the above quantities were computed for $a = z$ and $n = 2, 3, 4, 6, 8, 16, 32, \infty$. For the case of $a = 2$, we find that

$$K = [3\ell - 4(1 - 2^{-\ell})]^{-1}$$

and that as $\ell \rightarrow \infty$,

$$H_L, H_D, H \rightarrow \frac{4}{3} + \log_2 3 = 2.918 \text{ bits.}$$

The results are shown in the Table I and also are plotted in Fig. 9.

While H_0 and H_1 increase without limit as ℓ is increased, H_L , H_D , and H quickly approach a definite limit. This limit exists because we assumed that the decrease in joint probability as a function of *number*

TABLE I

Number of levels	H_0	H_1	H_L	H_D	H
1	0	0	0	0	0
2	1	1	1.252	0.918	0.918
3	1.585	1.583	1.777	1.437	1.422
4	2	1.995	2.074	1.764	1.750
6	2.583	2.575	2.381	2.157	2.131
8	3	2.988	2.552	2.370	2.343
16	4	3.99	2.768	2.678	2.641
32	5	5 - ϵ	2.850	2.818	2.782
∞	$\log \infty$	$\log \infty$	2.918	2.918	2.918

(These figures were computed by slide rule so the fourth figure is not very significant.)

of levels off the diagonal was the same regardless of ℓ . In typical signals this is not true. The decrease is more apt to depend on *amplitude difference* and the finer the quantum step, the more levels a given difference represents. As a result, the probability will fall off less per level off the diagonal, and doubling ℓ will in general add one bit to H .

On the other hand, doubling the sampling rate will not in general double the required channel capacity, for the closer spaced samples will

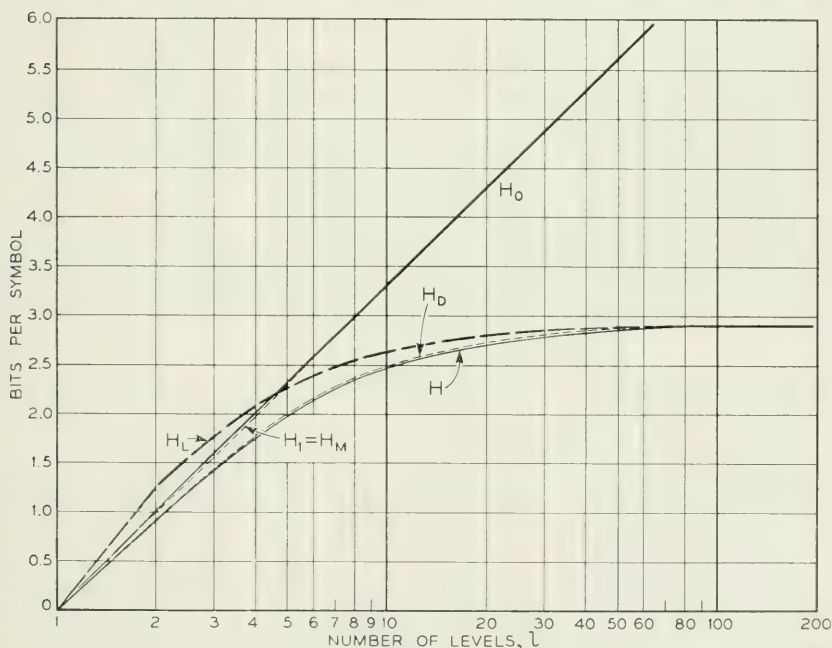


Fig. 9.

be more highly correlated. Thus in TV, doubling the horizontal resolution would not double the bandwidth for the same picture material if use were made of the statistics. (Of course, increased resolution in TV might encourage the use of more detailed scenes and this *would* increase the required bandwidth.)

It should also be noticed that for small ℓ , linear prediction actually makes matters worse. The increase in the number of levels in the error signal more than offsets the peaking of the distribution.

Since all the conditional distributions in this message are of similar shape (after ordering), H_D and H are almost the same, for all ℓ . The difference between H_L and H_D is slight except for small ℓ because the distribution we assumed is unimodal throughout.

Fig. 10 shows the simple probability distributions for (a) the original message, (b) the reduced signal from linear prediction, and (c) the reduced signal from the digrammer.

VARIABLE DELAY AND OTHER PROBLEMS

We have seen in the last two sections how it is possible to convert a message for which $H \ll H_0$ as a result primarily of intersymbol correlation, into a reduced signal for which $H \ll H_0$ as a result primarily of a highly peaked probability distribution in the individual symbols (i.e. one for which $H_1 \rightarrow H$). Since the operations are reversible, the true information rate, H , is preserved. In the original signal it was the *conditional* distributions which were peaked, while the simple distribution was relatively flat. In the reduced signal the *simple* distribution is peaked.

The result is that whereas a Shannon-Fano code would only have been effective on the original message if applied to blocks two or more symbols in length (and then it would ignore correlation between blocks), in the reduced signal the code will be effective on a symbol to symbol basis.

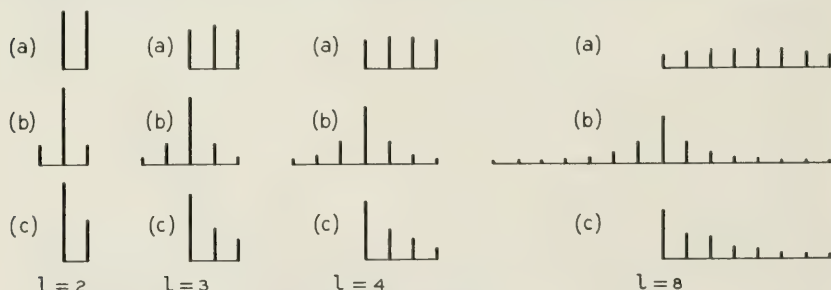


Fig. 10—Probability distributions: (a) Original, (b) After linear modal prediction, (c) After digramming.

The encoding of the reduced signal into binary digits presents no theoretical difficulties. A PCM type coder tube⁴ with the appropriate Shannon-Fano groups built into it is all that is needed. The biggest practical complication arises out of the fact that the code groups are of different length. Some messages, such as written text, can be fed into the system as fast as it can handle them. The transmission time will then vary with the message complexity. Others, such as television are generated and must be accepted and delivered at a constant rate. One solution is then to take the binary digits in big and little batches as they come from the coder and store the surplus in a sort of pulse "surge tank" before they are sent over the channel at a regular rate. At the receiver, a similar sort of storage register is necessary as the pulses arrive over the channel at a regular rate and are used by the decoder at a varying rate. Devices which will perform this variable delay function satisfactorily for signals with relatively slow sampling frequency are available, and as the art progresses there is every reason to believe that high speed sampled signals like television can be handled also.

It will be noticed that the digramming or prediction operation, while it involves memory, does not introduce appreciable transmission delay. Each symbol of the reduced signal appears the moment the corresponding message sample is applied. The total transmission delay required for statistical coding thus depends upon how much variation is required in the variable delay units. This in turn depends upon the degree of stationarity in the "local information rate" of the message. For example, in television, if each line could be described (by the n -grammer and subsequent coder) in the same total number of binary digits, then the total delay variation and total delay would be less than one line time. Since this is not true, we either must have enough channel capacity to send in one line time the number of digits corresponding to the "worst" line, or enough variable delay to average the existing rate over many lines.

Probably the most practical solution is to provide sufficient channel capacity and variable delay to take care of all but a small fraction of the possible message sequences. Then when an unusual stretch of message continues long enough for the variable delay to be nearly all used up, the system should fail in some relatively harmless way. In television, the sampling rate could be momentarily reduced, for example. This would degrade the resolution in rare situations, but a small amount of this could be tolerated in return for transmission savings.

If long blocks of the message are efficiently encoded as a group, then an error in transmission may cause the whole block to be reproduced

incorrectly. If n -gramming or prediction is used, then an error in transmission will cause the receiver to function improperly not only for that symbol, but its further n -gram decoding or prediction will also be disturbed. Thus errors of transmission are either spread over definite blocks, or propagate for a considerable time rather than being confined to the particular symbols sent in error. In fact, if the encoding were completely efficient, all received sequences would be possible messages, and a single error could convert the received message from the proper one into a completely different but possible one. With *no* redundancy there is no way to recognize an error. It is for these reasons that we have assumed a rugged (quantized) channel. In view of the eight to ten db more average power required in a quantized channel to achieve the same channel capacity as an ideal channel of the same bandwidth, considerable statistical saving must be possible before statistical coding may be warranted. This initial handicap of course does not apply to channels already designed to work on a digital basis for other reasons. Lastly, the use of error correcting codes⁵ is a possibility. In these codes a small amount of redundancy is introduced in a particularly efficient fashion. As a result, a certain frequency of transmission errors can be tolerated without causing errors in the reproduced message.

REFERENCES

1. Oliver, Pierce, and Shannon, "The Philosophy of PCM," *Proc. Inst. Radio Engrs.*, Nov. 1948.
2. C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Tech. J.*, July and Oct. 1948; Shannon and Weaver, "The Mathematical Theory of Communication," University of Illinois Press, 1949.
3. C. E. Shannon, "Prediction and the Entropy of Printed English," *Bell System Tech. J.*, Jan. 1951.
4. R. W. Sears, "Electron Beam Deflection Tube for Pulse Code Modulation," *Bell System Tech. J.*, Jan. 1948; W. M. Goodall, "Television by Pulse Code Modulation," *Bell System Tech. J.*, Jan. 1951.
5. R. W. Hamming, "Error Detecting and Error Correcting Codes," *Bell System Tech. J.*, Apr. 1950.

Statistics of Television Signals

By E. R. KRETZMER

(Manuscript received February 28, 1952)

Measurements have been made of some basic statistical quantities characterizing picture signals. These include various amplitude distributions, autocorrelation, and correlation among successive frames. The methods of measurement are described, and the results are used to estimate the amount by which the channel capacity required for television transmission may be reduced through exploitation of the statistics measured.

INTRODUCTION

One of the teachings of information theory is that most communication signals convey information at a rate well below the capacity of the channels provided for them. The excess capacity is required to accommodate the redundancy, or repeated information, which the signals contain in addition to the actual information. Removal of some of this redundancy would reduce the channel capacity required for transmission, thus opening the way for possible bandwidth reduction. In order to remove redundancy, one must first understand it; the amount and nature of the redundancy can be completely defined in terms of various statistical parameters characterizing the signal.

It has been pointed out that the existence of redundancy is particularly evident in the case of television; moreover, its elimination is highly desirable because of the large bandwidth presently required for transmission. Evidence of redundancy is found in the subject matter of television—the average scene or picture. Knowing part of a picture, one can generally draw certain inferences about the remainder; or, knowing a sequence of frames, one can, on the average, make a good guess or prediction about the next frame. In either case, knowledge of the past removes uncertainty as to the future, leaving less actual information to be transmitted.

Another way of looking at this is to visualize the picture as an array of approximately 210,000 dots, 500 vertically, 420 horizontally, corresponding, respectively, to the 500 scanning lines and 420 resolvable

picture elements per line of the standard television raster. Each dot can have, say, 100 distinguishable brightness values in a good-quality picture. The number of possible combinations is therefore approximately $100^{210,000}$ or $10^{420,000}$. At the usual rate of 30 frames per second it would take approximately $10^{419,991}$ years to transmit all these "pictures," which our present television system is fully prepared to transmit! The vast majority of these "pictures" will, of course, never be transmitted in this age because the average picture statistics virtually preclude the possibility of their occurrence.

If all of the redundancy alluded to in the preceding paragraph were to be expressed in terms of statistics, the array of data would be staggering.* Redundancy encompassing even a small part of a single frame implies statistics of enormously high order because of the large number of possible past histories. The initial attention should therefore be focused on local redundancy, encompassing only a few adjoining picture elements. Accordingly, measurements have been made of the following statistical quantities.

1. *Simple probability distribution of signal amplitudes corresponding to picture brightness.* This encompasses only a single picture element, revealing the relative probabilities of this or any element's assuming the various possible brightness values, in the absence of any past-history information.

2. *Simple probability distribution of error amplitudes resulting from linear prediction of television signals.* Only the simplest type of linear prediction is considered here, so-called previous-value prediction, which predicts each picture element to have the same brightness value as the preceding one. The prediction error signal is simply the difference between the picture signal and a replica delayed by one Nyquist interval (one-half the reciprocal bandwidth or the time interval corresponding to the spacing between picture elements). The distribution of this error signal encompasses two picture elements (past history of one element) and therefore is a condensed version of the family of first-order joint probability distributions.

3. *Autocorrelation of typical pictures.* This statistical quantity is an even more streamlined version of various families of different-order joint probability distributions. Each family corresponds to just a single point on the autocorrelation curve; the ordinates of the curve represent the average correlation between picture elements spaced by various

* Complete statistics extending, say, over one frame period, would comprise one conditional probability distribution per picture element for each possible past history. With the approximate figures cited above, the number of distribution curves (many of which would be similar) is $210,000 \times 10^{419,999}$ or $10^{420,004.3}$.

distances. This correlation, say, between horizontally adjoining elements is simply the average product of the two brightness values of each pair of neighbors, relative to the average square of all brightness values.

The three quantities enumerated above contain a great deal of statistics in very compact form, but these statistics are essentially of a local and linear nature. They do not include the bulk of the large-scale redundancy, which is of a far-flung and nonlinear nature.

AUTOCORRELATION

For a function of time, $f(t)$, the autocorrelation can be expressed as

$$\phi(\tau) = \overline{f(t) f(t + \tau)} \quad (1)$$

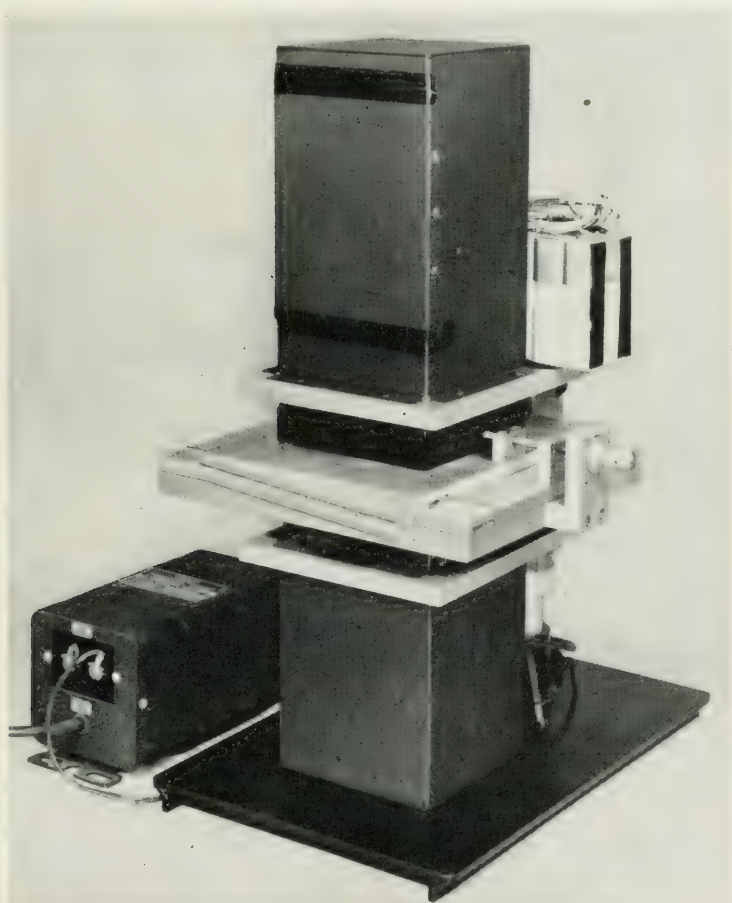


Fig. 1—Picture autocorrelator.

averaged over all time, for various values of the time shift τ . In the case of a picture transparency, the optical transmission is a function of two-dimensional space, expressible in polar coordinates as $T(s/\phi)$, and the autocorrelation can be expressed in analogous fashion. The time variable t is replaced by the space coordinate s/ϕ , and the correlation time shift τ is replaced by a space shift $\Delta s/\theta$, so that the new expression is

$$\phi(\Delta s/\theta) = \overline{T(s/\theta) T(s/\theta + \Delta s/\theta)}, \quad (2)$$

averaged over as much area as practicable. This space-domain autocorrelation is much easier to measure than the time-domain autocorrelation. We need merely measure the relative optical transmission of two identical cascaded transparencies, shifted from register by a variable

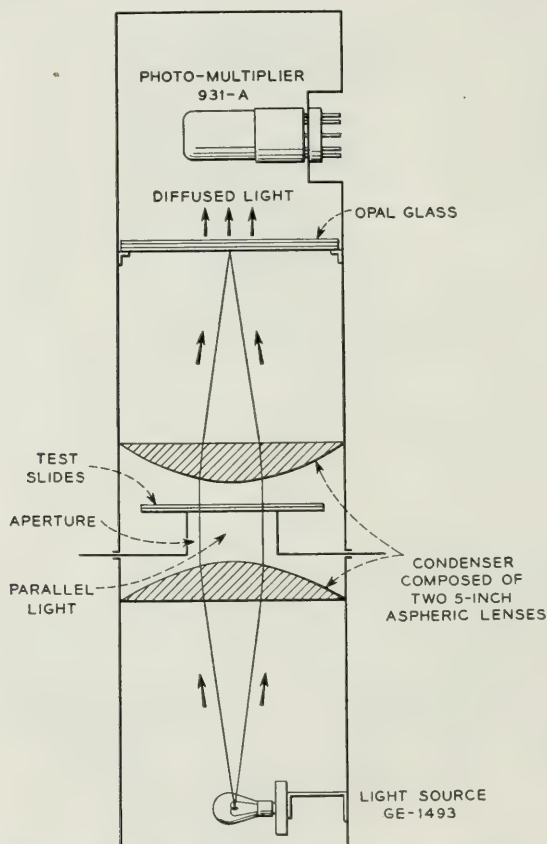


Fig. 2—Basic arrangement of picture autocorrelator.

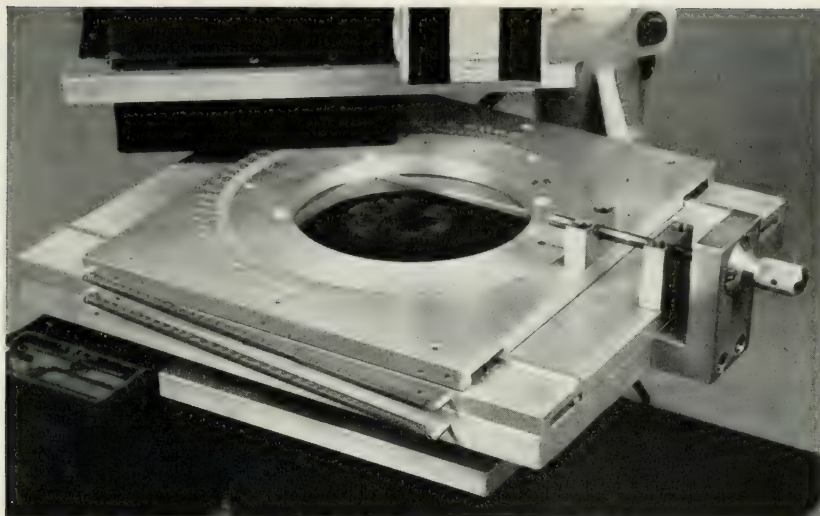


Fig. 3—Close-up view of slide holding assembly and shifting mechanism of picture autocorrelator.

amount. The averaging process is inherent in such a measurement.

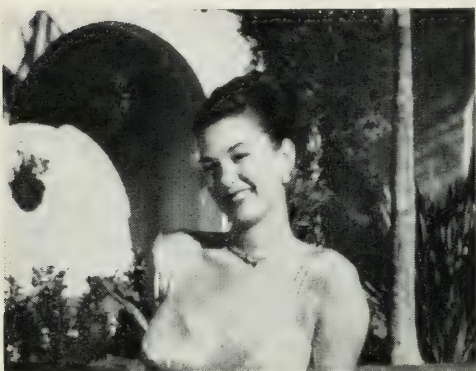
The apparatus used to measure autocorrelation is shown in Figs. 1 and 2. The chamber at the bottom contains a light source of very constant intensity and a convex lens to collimate the light. The middle part, made of accurately machined aluminum, holds the two identical slides of the picture under test, and an aperture exposing a large circular area of the slides. The top chamber contains a collector lens and a photomultiplier tube which (on a microammeter not shown) gives a sensitive indication of the total light transmitted through the slides. Fig. 3 shows a close-up view of the slide-holding assembly. Two close-fitting graduated aluminum rings permit accurately determined rotation of both slides or one slide, and the micrometer drive permits translational displacements measurable to within one mil (moving the two slides by equal and opposite amounts); the separation between picture elements is approximately 7.5 mils horizontally and 5 mils vertically (for the $2\frac{1}{2}$ " by $3\frac{1}{4}$ " slide size used).

The light transmission is always a maximum when the two slides are in precise register ($\Delta s = 0$). For large shifts the transmission fluctuates about a nonzero asymptote. The nonzero asymptote results from the fact that the average transmission is always positive, and the fluctuation from the fact that large displacements introduce substantial amounts of new picture material into the aperture. Since these components tend to

obscure the correlation effects, it is useful to make additional measurements which enable us to subtract them out completely. This leaves us with a 'pure' autocorrelation $A(\Delta s/\theta)$, which is then normalized so as to have a peak value of unity. It is given by

$$A(\Delta s/\theta) = \frac{T_2\left(\pm \frac{\Delta s}{2} / \theta\right) - T_1\left(\frac{\Delta s}{2} / \theta\right) T_1\left(-\frac{\Delta s}{2} / \theta\right)}{T_2(0) - T_1^2(0)}, \quad (3)$$

where $T_2\left(\pm \frac{\Delta s}{2} / \theta\right)$ is the transmission through the two cascaded slides shifted by equal and opposite amounts $\frac{\Delta s}{2}$ at an angle θ with the horizontal, and $T_1\left(\frac{\Delta s}{2} / \theta\right)$ is the transmission of a single slide with displacement $\frac{\Delta s}{2}$ at the same angle θ .



SCENE A



SCENE B



SCENE C



SCENE D

Fig. 4—Test pictures whose statistics are included in this article.

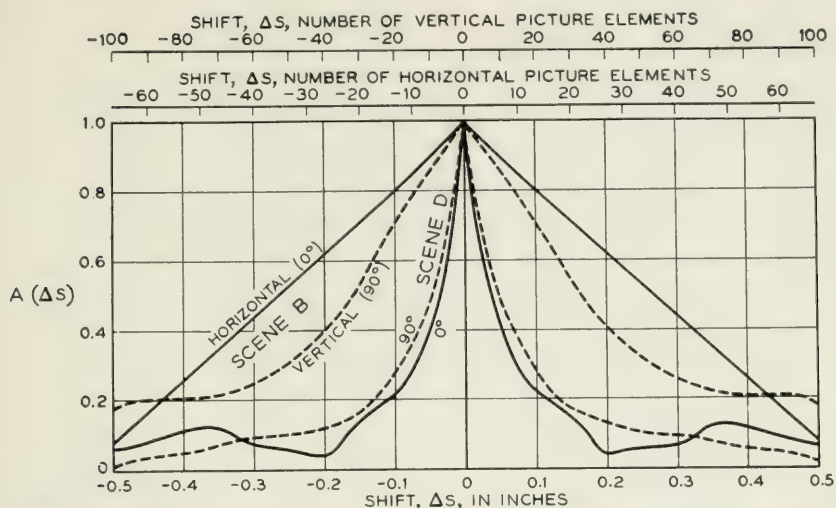


Fig. 5—Plots of autocorrelation in horizontal and vertical directions for two pictures.

Fig. 4 shows some pictures for which autocorrelation measurements have been made. The results can be presented in the various ways shown in Figs. 5, 6, and 7. Fig. 5 shows conventional plots of A versus Δs in the horizontal and vertical directions. Scene B is seen to have more correlation than Scene D, and curve shapes range from remarkably linear to somewhat like exponential. Fig. 6, giving contours of constant autocorrelation, brings out the variation with the angle θ . Scene A happens to have its greatest correlation in the vertical direction, but that was not found to be a general rule by any means; Scene B, for example, has its greatest correlation in the horizontal direction. No preferred directions appear to exist in general. In Fig. 7 attention is focused on the more local correlation, for small values of Δs . The average correlation among horizontally adjoining picture elements, designated by A_{10} , is seen to be approximately 0.99 for Scene B and only 0.75 for Scene C. A_{20} denotes the correlation for a horizontal spacing of two picture elements while A_{01} denotes the correlation among vertically adjoining picture elements.

It should be pointed out that the pictures which gave the above results were not band-limited to the standard 4-mc resolution. However, before the results were used quantitatively, the proper band limitation was applied mathematically. This has the effect of rounding off the peaks of the curves, decreasing the autocorrelation drop within the first Nyquist interval by up to approximately 24 per cent.

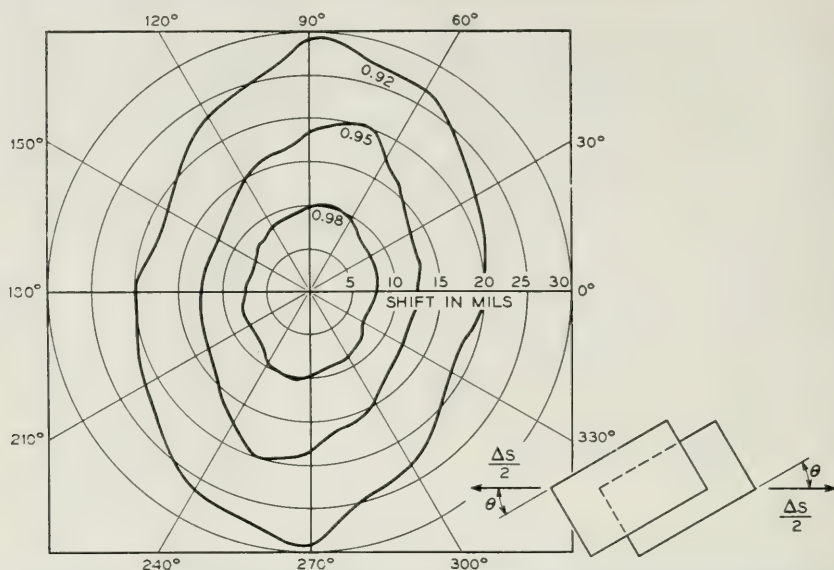


Fig. 6—Contours of constant autocorrelation for Scene A. In general there are no preferred directions of correlation.

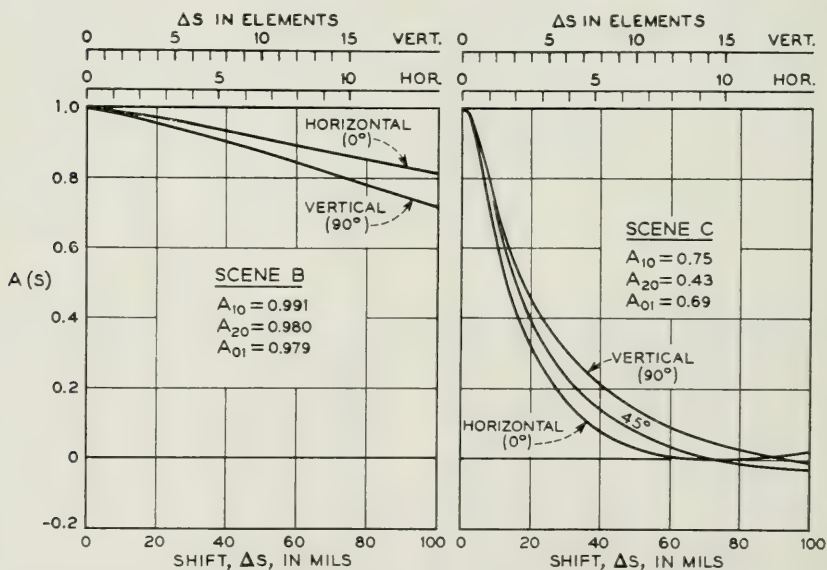


Fig. 7—Plots of autocorrelation for small shifts. A_{10} is the autocorrelation for a shift of one horizontal elemental distance, A_{20} for two horizontal elemental distances, and A_{01} for one vertical elemental distance. Alternatively A_{10} may be described as the average correlation between horizontally adjoining elements, etc.

PROBABILITY DISTRIBUTIONS

A probability distribution of amplitudes is generally shown as a plot of probability density versus signal amplitude. Probability density, say, corresponding to amplitude x_1 , is the probability of finding the signal amplitude between x_1 and $x_1 + dx$, divided by the differential amplitude increment dx . Conversely, the probability of finding the signal amplitude between x_1 and $x_1 + dx$ is given by $p(x_1)dx$, $p(x)$ being the probability density corresponding to amplitude x .

If a cathode-ray spot is deflected, say horizontally, by the signal in question, its average dwell time at any point is directly proportional to the corresponding probability density. In the optical system shown in Fig. 8, a cylindrical lens maps each point into a vertical line which is

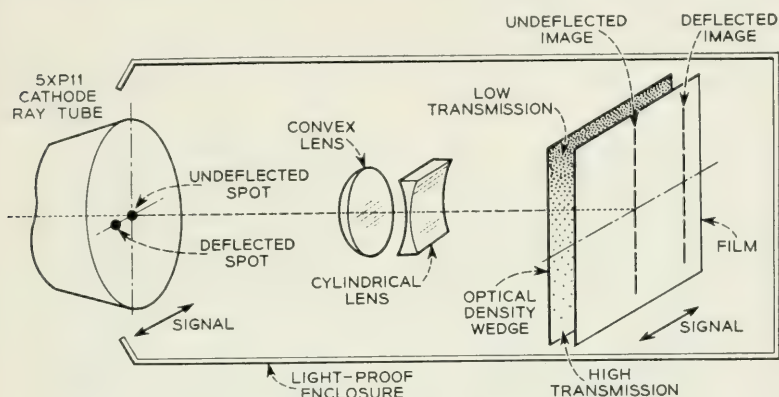


Fig. 8—Basic arrangement of probabloscope.

then tapered in intensity by an optical density wedge before reaching a high-contrast photographic film. Depending on the dwell time at any amplitude level, the corresponding tapered line has enough average intensity to blacken the film up to a certain level. This level is proportional to $\log p(x)$, since the density wedge is tapered exponentially so that the intensity of each tapered line of light reaching the film diminishes, say, by a factor of ten for each inch we travel up the line. The film in effect traces out a contour of constant exposure.

Two or three iterated photographic printings increase the effective gamma sufficiently to yield a contour of ample sharpness. This contour is then changed to a sharp line by a simple dark room trick: while the film is in the development tray, already fully developed, it is momentarily exposed to light. The blackened portion of the film is unaffected, the clear portion is fully blackened, while the transition contour, being partly opaque, is not fully blackened. By printing from this film we then

obtain a well-defined black-on-white curve of $p(x)$ versus x on a logarithmic probability scale. The logarithmic scale has the advantage of making the curve shape independent of exposure length and giving uniform relative accuracy over the entire range.

Fig. 9 shows some typical results obtained by means of the "probabiloscope." The two small curves are distributions of two different still pictures. The left-hand end corresponds to black, the right-hand end to peak white; the blanking intervals (slightly blacker than black) cause the peaks at the extreme left. (The signals did not contain any synchronizing pulses.) The tall and slender curve at the right of Fig. 9 is the distribution of errors resulting from previous-value prediction of one of the pictures in Fig. 4. The peak corresponds to zero error which is seen to be most probable, as it should be if the prediction criterion is good. Increasingly larger errors are increasingly improbable or rare. The six decades of probability density spanned by the curve were obtained in three separate exposures and subsequently joined, since stray

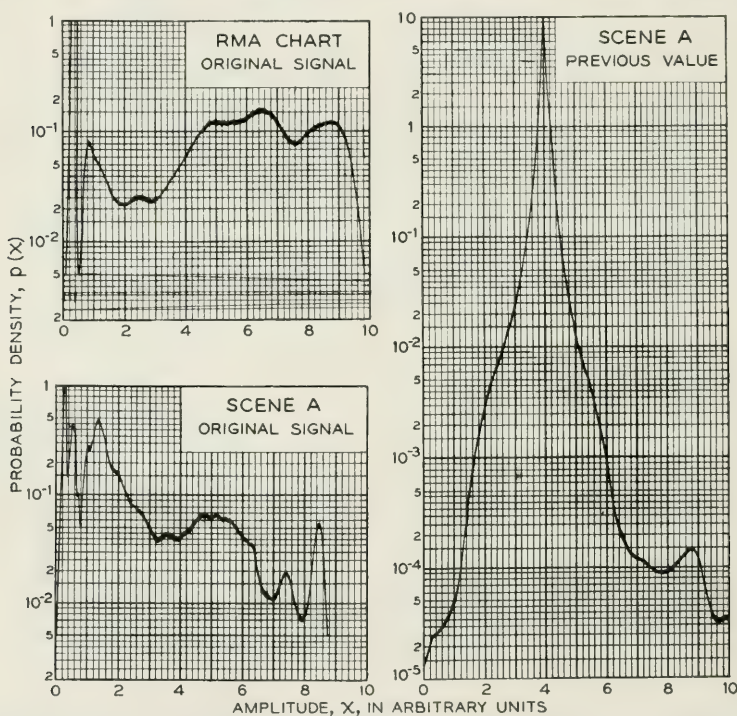


Fig. 9—Typical probability distributions as obtained from the probabiloscope. Curves at left are for video signals; right-hand curve is for difference between video signal and delayed replica.

light limits the useful range of the probabiloscope to approximately two decades. In obtaining those sections of the curve corresponding to the few and far-between large errors, a long exposure was used and the cathode-ray beam was blanked whenever passing through the range of zero or small errors. The vertical scale on all curves is determined solely by the density taper of the optical density wedge. If this scale is to represent true probability density, instead of a proportional quantity, it should be shifted up or down so as to make the area under the curve equal to unity.

APPLICATION OF RESULTS

The statistics measured can be put to various uses, such as in the design of better predicting or coding schemes. The most interesting application is probably in estimating the reduction in channel capacity which the measured statistics show to be theoretically possible. In other words, the results can give us various lower bounds to the redundancy of television signals.

For the sake of illustration, suppose that the signal is quantized into 64 amplitude levels. An ordinary television channel assumes all 64 levels to be equally likely, hence is prepared to accommodate $\log_2 64$ or 6 bits per sample. But the simple amplitude distribution of the signal is not flat, so that all 64 levels are *not* equally likely. The maximum possible associated average information content per sample is given by

$$H_{\max} = \sum_i^{64} p_i \log p_i, \quad (4)$$

where p_i is the simple probability of the signal's falling into the i th level. Since the 64 p_i 's are unequal, H_{\max} is necessarily less than 6 bits. For all available data the average value of H_{\max} turns out to be approximately 5 bits, indicating a one-bit redundancy. The latter figure is essentially independent of quantization.

The prediction error signal still contains all the useful picture information. The maximum possible information content per sample (maximum in that all samples are assumed to be completely independent) is still given by (4) but in this case the 64 values of p_i are obtained from the peaked error distribution. The average* result from all available data turns out to be approximately 3.4 bits below the 6-bit ceiling, show-

* This average was computed by averaging the various redundancy values obtained for the individual pictures, rather than averaging all statistical data and then finding one corresponding average redundancy. The average computed here is more favorable and can be realized only if optimum coding is performed on a short-term basis rather than on the basis of one set of long-term statistics.

ing that the original signal must have contained at least 3.4 bits of redundancy.

The autocorrelation can also furnish a lower bound to the redundancy, as has been pointed out by P. Elias in his Letter to the Editor of the *Proceedings of the I.R.E.* for July, 1951. If, for example, the correlation A_{10} , between horizontally adjoining picture elements, is high, the corresponding lower-bound redundancy is very roughly equal to

$$R \approx -\frac{1}{2} \log_2 (1 - A_{10}) \text{ bits/sample.} \quad (5)$$

Alternatively, taking the Fourier transform of the autocorrelation yields the power spectrum $P(f)$, from which we can find the lower-bound redundancy through the relation

$$R = \frac{1}{2W} \int_0^W \log_2 P(f) df + \frac{1}{2} \log_2 W + \log_2 K \text{ bits/sample,} \quad (6)$$

where W = bandwidth in cps, and $\frac{1}{K} = \int_0^W P(f) df$.

Using either method, one obtains approximately 2.4 bits for the average* of the available data. This is an approximate bound, in that it applies strictly only to functions having gaussian amplitude distributions.

Suppose, then, that we have exposed an average redundancy of at least 3 bits per sample. This means a potential 3-bit reduction in the channel capacity required for television transmission. In a 6-bit system (64 amplitude levels) this means a 50 per cent reduction, and hence a potential halving of the bandwidth with the aid of an ideal coding scheme. It is true that the decorrelated signal is somewhat "frail," i.e., vulnerable to interference, so that it might be desirable to use a "rugged" system of the PCM variety for transmission. Thus, if a Shannon-Fano code were used, the 3-bit decorrelation should enable us to send television by an average of 3 on-off pulses per picture sample rather than 6. This represents a two-to-one saving over the usual PCM bandwidth. More spectacular reductions are likely to be achievable only by tapping the large-scale redundancies mentioned earlier.

FRAME-TO-FRAME CORRELATION

There is, of course, a great deal of interest in the possibility of utilizing the similarity between successive frames. Accordingly, adjacent-frame

* See previous footnote.

correlation was measured for two typical motion-picture films, by means of the apparatus described in the section on autocorrelation.* The results were 0.80 and 0.86, after correction for the 4-mc bandwidth limitation. This means that "previous-frame" prediction can remove only slightly more than one bit of redundancy per sample. More complicated schemes would presumably be more successful in taking advantage of the large frame-to-frame redundancy which undoubtedly exists.

ACKNOWLEDGMENT

Many of the ideas expressed in this paper are due to B. M. Oliver, whose resourcefulness is hereby gratefully acknowledged.

* The expression used in evaluating the correlation between frame 1 and frame 2 (any two frames) is

$$C_{12} = \frac{T_{12} - T_1^2}{T_{11} - T_1^2}, \quad (7)$$

where T_{12} is the optical transmission of frames 1 and 2 in cascade, T_1 is the average of the individual transmission of frames 1 and 2, and T_{11} is the average of the transmissions of two cascaded slides of frame 1 and two cascaded slides of frame 2, respectively. In all cascade transmission measurements, the two frames must be in precise register.

Experiments with Linear Prediction in Television

By C. W. HARRISON

(Manuscript received February 28, 1952)

The correlation present in a signal makes possible the prediction of the future of the signal in terms of the past and present. If the method used for prediction makes full use of the entire pertinent past, then the error signal—the difference between the actual and the predicted signal—will be a completely random wave of lower power than the original signal but containing all the information of the original.

One method of prediction, which does not make full use of the past, but which is nevertheless remarkably effective with certain signals and also appealing because of its relative simplicity, is linear prediction. Here the prediction for the next signal sample is simply the sum of previous signal samples each multiplied by an appropriate weighting factor. The best values for the weighting coefficients depend upon the statistics of the signal, but once they have been determined the prediction may be done with relatively simple apparatus.

This paper describes the apparatus used for some experiments on linear prediction of television signals, and describes the results obtained to date.

INTRODUCTION

Linear prediction is perhaps the most expedient elementary means of removing first order correlation in a television message. Before discussing the advantages and disadvantages of linear prediction, it might be well to consider what is generally meant by correlation in a television picture and why it should be removed.

Almost every picture that has recognizable features contains both linear and non-linear correlation. Each type of correlation helps in identifying one picture from another; however, linear prediction is only effective in removing linear correlation, and for this reason, future references to correlation will refer only to its linear properties. With television, a signal is obtained as the result of scanning; hence, the cor-

relation is evident in both space and time. Briefly, correlation is that relation which the "next" elemental part of the signal has with its past.

To leave correlation in a message is to be redundant, and this effectively loads the transmission medium with a lot of excess "words" not necessary to the description of the picture at the receiving end. It is then more "efficient" to send only the information necessary to identify the picture, and to restore the redundancy at the receiver.

EFFICIENT TRANSMISSION

The more *efficient* we are in sending pictures over a given transmission line, the more alarmed we become at the increasing amount of equipment that is required at the transmitting and receiving terminals. Certainly the design will be a compromise between the complexity of apparatus and the efficiency achieved. The ingenuity of engineers will be taxed along these lines for years to come; however basically, the general form of these systems will be similar to that shown in Fig. 1. Although not always separable, four essential operations are required—namely, decorrelating, encoding, decoding and correlating. The transmitting decorrelator and the encoder encompass the principal design problems, since the decoder and correlator at the receiving end perform the reverse operations which interpret the code and add in the redundancy that was removed.

Decorrelation involves prediction, and as the predictors are more nearly made to predict the future of the signal, the more the output signal from the decorrelator resembles random noise. The essential picture information is still present, which means that our original picture signal can be obtained at the receiving end without theoretical degradation. The basic job of the encoder is to match the picture information out of the decorrelator to the channel over which it is to be transmitted. There are several encoding operations. The first concerns the rate of information into the encoder, and that required out of it. In the case of television, there are flat, highly correlated areas as well as areas containing more concentrated detail. This means that the information

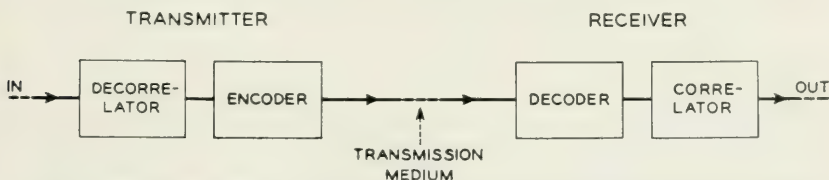


Fig. 1—Block diagram of an efficient transmission system employing reversible decorrelating and encoding means.

rate varies when the picture is scanned at the conventional uniform scanning rate. The output of the encoder feeds a transmission line that has a definite channel capacity, and if maximum efficiency is to be obtained from this transmission medium, then the rate of information into it must be held relatively uniform at a value near the channel capacity. It is the job of the encoder to take the varying rate of information from the decorrelator and feed it to the channel at a constant rate. At the receiving end, the decoder must take the constant rate of information and deliver it to the correlator at the variable rate as originally fed into the transmitter's encoder. Thus, to perform this task, a variable or elastic delay to run ahead or behind, depending on the information content of the picture being scanned, is an important part of the encoder. Over a long period of time, the variable delay would average out to some fixed value. This variable delay must never run out, even when the detail is concentrated. There are instances when this condition could not be met, such as an extended reproduction of a snow storm; however, with good design the system should fail "safe"—a slight degradation of picture quality. This condition can be made infrequent enough to cause little concern.

The encoder design must also account for noise as well as bandwidth of the channel and must consider the ultimate effect of an error that may be introduced by noise along the transmission line. As more redundancy is removed to get at the "essence" of the picture signal, the more important it is to guard this "essence," as mistakes presented to the receiver will propagate themselves longer in the absence of correlation. Errors can be minimized by rugged systems of modulation such as PCM, where the signal-to-noise ratio of the transmission line determines the base of the PCM system selected. In any event, the encoder must send the information so that the effect of errors will not appreciably disturb the picture.

DECORRELATION AND LINEAR PREDICTION

Fig. 2 illustrates, in a general way, a means of decorrelating the signal, $S_1(t)$. For purposes of explanation, the encoder and decoder have been omitted, and the transmission between the receiving and sending terminals, idealized. The predictors, P , are identical, and base their prediction, $S_p(t)$, on the signal's past history. In this way, the output of the computer represents the discrepancy between the actual value of the signal sample and the predictor's prediction. By this means we are sending only our mistakes—the amount by which the next picture element surprises us. For example, if the computer is so designed that it

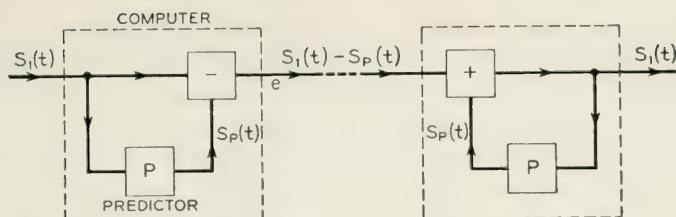


Fig. 2—Decorrelator and correlator showing reversible nature of this method of removing redundancy.

bases its prediction on the “previous frame,” and we are transmitting a “still,” there will be no surprises after the first frame and consequently no output signal. Certainly it is redundant to send the same picture more than once.

Linear prediction provides an easily instrumented means of removing redundancy. With linear prediction the next signal sample is simply the sum of the previous signal samples, each multiplied by an appropriate weighting factor. The best values for these weighting coefficients depend on the statistics of the signal.

Fig. 3 is a block diagram of a decorrelator employing linear prediction. The delayed versions of the input signal can be obtained from taps along the delay line. The weighting coefficients for each of the delayed signals are selected by loss in their respective paths as shown by the amplitude controls. The polarity of each signal can be determined by the switches. The output is simply the sum of these weighted signals.

If we consider the signal on a continuous basis (not quantized or sampled), linear predictors can be characterized as ordinary linear filters used to predistort the frequency spectrum of the signal. As such, they can be designed in the frequency or time domain. However as will

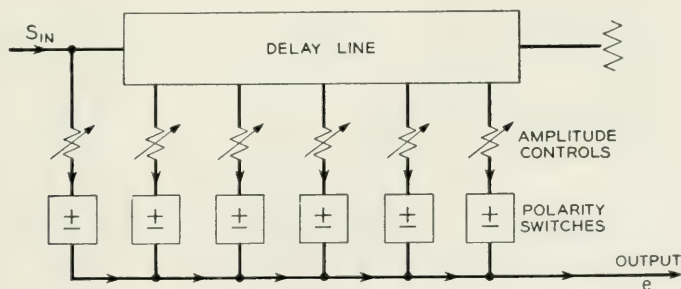


Fig. 3—General block diagram of decorrelator employing linear prediction. Linear prediction bases its prediction on the weighted sum of previous signal samples.

be shown, it is much easier to recognize circuit configurations that reduce redundancy in the time domain. To this end, and for purposes of encoding, the signal is thought of as signal samples uniformly spaced at Nyquist intervals. Thus, amplitude values obtained by sampling a 4.0 mc picture signal at $\frac{1}{8}$ microsecond intervals serve to specify the signal completely. Fig. 4 shows a small portion of a television raster where the signal is represented by signal values spaced at Nyquist intervals, τ . The coordinates shown are designated with respect to the "present value" of the signal, $S_{0,0}$. The positive coordinate directions are shown by the arrows. The past is represented by positive coordinates—the future by negative coordinates. In this way, the previous value of the signal

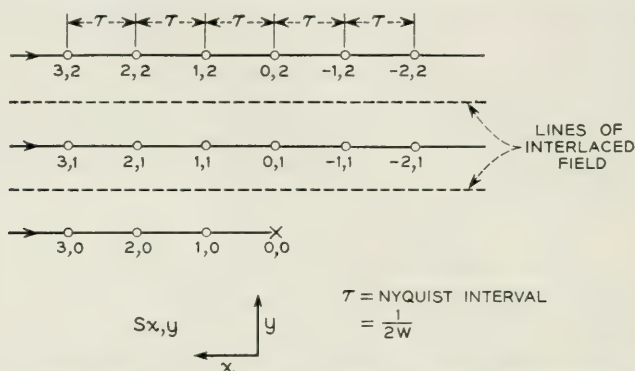


Fig. 4—A small portion of a television raster showing geometrical location of signal samples with relation to the "present value" of the signal, $S_{0,0}$.

taken one Nyquist interval before $S_{0,0}$ is designated by $S_{1,0}$ —the previous line samples by $S_{0,1}$, etc.

METHODS OF LINEAR PREDICTION

As previously stated, with linear prediction the next signal sample, $S_p(t)$, is simply the sum of the previous signal samples each multiplied by an appropriate weighting factor. Thus,

$$S_p(t) = a_{1,0}S_{1,0} + a_{2,0}S_{2,0} + a_{3,0}S_{3,0} + \cdots a_{m,n}S_{m,n}$$

represents the weighted sum of all the previous signal values. The error signal, e , as shown in Fig. 2, is represented by the difference between the present value of the signal, $S_{0,0}$ and the predictor's prediction.

$$e = S_{0,0} - S_p(t)$$

There are several specific types of linear prediction that deserve fur-

ther explanation—namely, “previous value,” “slope,” “previous line,” “planar” and “circular.”

“*Previous value*” prediction is illustrated in Fig. 5. Here the prediction is taken to be the signal amplitude of the preceding picture element. The previous amplitude of the signal, $S_{1,0}$ is subtracted from the present value of the signal, $S_{0,0}$. The error signal is given as

$$e = S_{0,0} - S_{1,0}$$

The filter characteristic can be expressed as

$$F(\omega) = \left(2 \sin \frac{\omega\tau}{2} \right) \epsilon^{i\left(\frac{\pi}{2} - \frac{\omega\tau}{2}\right)}$$

This method of prediction proves to be rather effective in reducing the average power for most television pictures. The expression for the filter characteristic given above shows that the peak amplitude can be twice that of the original signal.

“*Slope*” prediction is illustrated by Fig. 6. “Slope” prediction is so called because it is equivalent to passing a straight line through the two previous signal values, with the assumption that this line will pass through the next signal value. The predicted signal is given by

$$S_p = 2S_{1,0} - S_{2,0}$$

The frequency and phase characteristic is expressed as

$$F(\omega) = [4 \sin^2 \omega\tau] \epsilon^{i(\pi - \omega\tau)}$$

For this method of prediction, the peak amplitude of the error signal can be as much as four times the peak amplitude of the original signal.

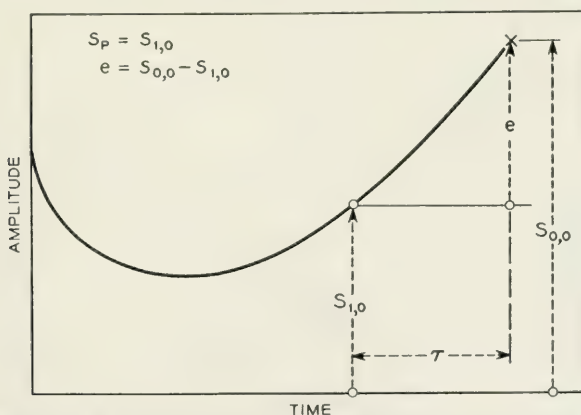


Fig. 5—Example of “previous value” prediction, where the error signal is the difference between the actual value of the signal and the previous value.

It is of interest to mention that "slope" prediction is equivalent to two "previous value" predictors in tandem. Three or more "previous value" predictors in tandem are equivalent to a binomial weighting of the previous values of the signal to form a predicted signal.

For example, the prediction for three "previous value" predictors in tandem is given by

$$S_p = 3S_{1,0} - 3S_{2,0} + S_{3,0}$$

For four "previous value" predictors in tandem

$$S_p = 4S_{1,0} - 6S_{2,0} + 4S_{3,0} - S_{4,0}$$

As can be seen from the above equations, further extension of "previous value" tandem operation results in a heavier weighting of picture elements further and further from the point to be predicted. For most pictures this leads to greater errors.

"Previous line" prediction, shown in Fig. 7, would be expected to be similar to previous value prediction, since a picture would presumably have approximately the same correlation vertically as it does horizontally. This would be the case except that our interlaced scanning system makes the previous line signal some 28 per cent further away from $S_{0,0}$ than the closest horizontal sample, $S_{1,0}$. The error signal, e , is given by where T is a line time. The error output has a maximum peak amplitude of twice the input.

$$e = S_{0,0} - S_{0,1}$$

The filter characteristic can be expressed as

$$F(\omega) = \left[2 \sin \frac{\omega T}{2} \right] e^{i\left(\frac{\pi}{2} - \frac{\omega T}{2}\right)}$$

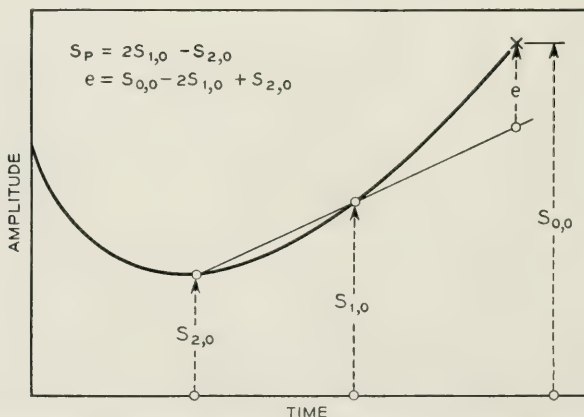


Fig. 6—Example of "slope" prediction. Here the next signal value is assumed to lie on a straight line that intersects the two previous signal values.

"Planar" prediction, shown in Fig. 8, is effectively tandem operation of "previous value" and "previous line" prediction. Planar prediction may also be thought of as the value represented by a plane above the present value of the signal when passed through three adjacent signal

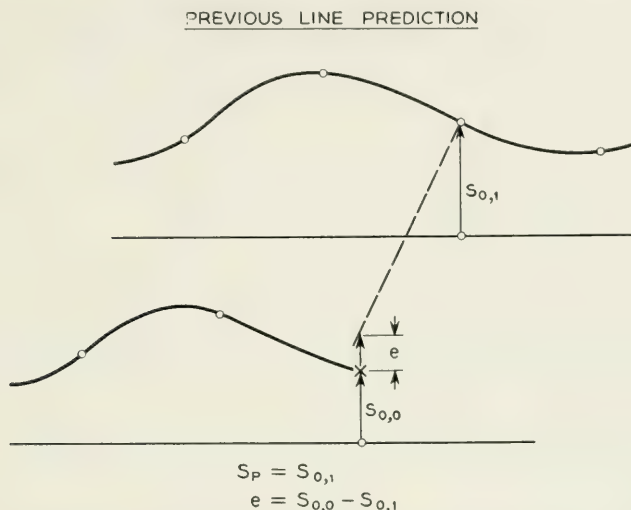


Fig. 7—Example of "previous line" prediction. Here the error signal is the difference between the actual value of the signal and the value of the signal on the line directly above.

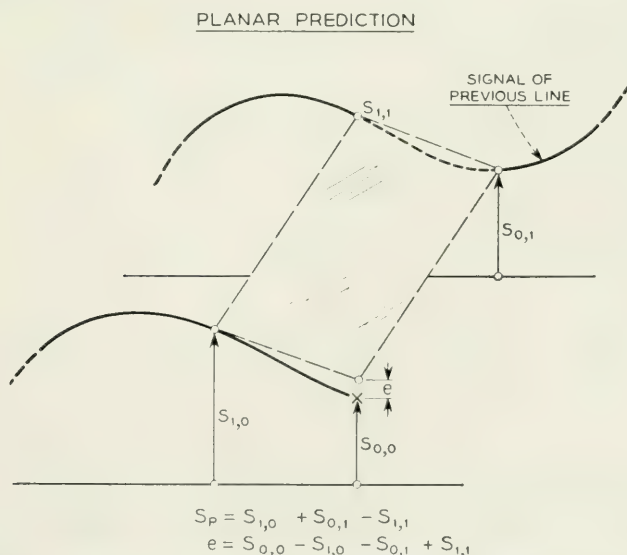


Fig. 8—An example of "planar" prediction. Here the prediction is represented by a plane that has been passed through three adjacent signal values.

values, namely $S_{1,0}$, $S_{0,1}$ and $S_{1,1}$. The predicted signal is given by

$$S_p = S_{1,0} + S_{0,1} - S_{1,1}.$$

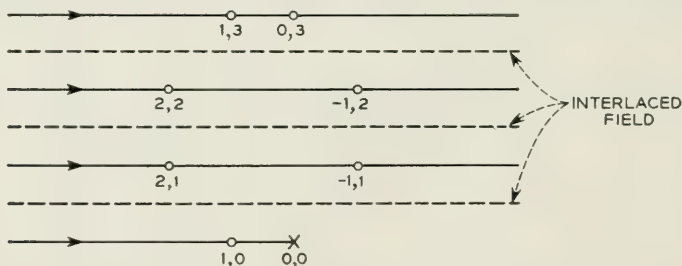
The filter characteristic is given by

$$F(\omega) = 4 \sin \frac{\omega\tau}{2} \sin \frac{\omega T}{2} e^{i\left(\pi - \frac{\omega}{2}(\tau + T)\right)}$$

The peak error amplitude for "Planar" prediction can be four times that of the input signal.

"Planar" prediction has several good characteristics. For example, if $S_{1,1}$ and $S_{1,0}$ were white and $S_{0,1}$ black, then $S_{0,0}$ would be predicted to be black. Thus a change horizontally from white to black would produce no errors. Similarly, if $S_{1,1}$ and $S_{0,1}$ were white and $S_{1,0}$ black, then $S_{0,0}$ would be predicted to be black. This indicates that a change vertically from white to black would be predicted correctly. In this manner, all vertical and horizontal contours in a picture are deleted. This philosophy can be extended to include other directions as well.

"Circular" prediction, illustrated in Fig. 9, is an extension of planar, since it deletes horizontal, vertical and 52° contours as well. A total of 190.5 microseconds of delay is required, making the required equipment more elaborate. Also, as more delay is required, more noise is added.



$$S_p = S_{1,0} - S_{2,1} + S_{2,2} - S_{1,3} + S_{0,3} - S_{-1,2} + S_{-1,1}$$



Fig. 9—Past signal samples required for "circular prediction"—a type of prediction which removes horizontal, vertical, and $\pm 52^\circ$ straight line picture contours.

Therefore, indefinite extension of this straight line contour deleting philosophy is not a paying means of prediction, at least not at the present state of the art of wide band delay lines. Furthermore, the increasing diameter of the circle for extension of circular prediction would decrease its accuracy for finely concentrated detail.

Fig. 10 shows the relative position of picture elements nearest $S_{0,0}$ if a wide band field delay were available. The methods of prediction dis-

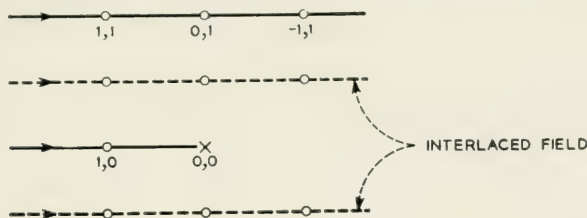


Fig. 10—Small portion of television raster showing signal samples, including those of previous field, which would enable time extrapolation-space interpolation as a method of prediction.

cussed have been essentially an extrapolation in space; however, with a field delay, interpolation in space, and extrapolation in time would also be possible.

EXPERIMENTAL CIRCUITRY

Experimentally, those types of predictors that involve only a few Nyquist intervals of delay are easiest to mechanize. Fig. 11 shows a simplified schematic of a decorrelator that enables an evaluation of linear prediction schemes having error signals given by $e = a_{0,0}S_{0,0} \pm a_{1,0}S_{1,0} \pm a_{2,0}S_{2,0}$. This enables an evaluation of "previous value" and "slope" prediction. The signal is fed into a terminated delay line having taps at Nyquist intervals. Each of these signals is individually attenuated by the potentiometers in the cathode circuit of the cathode followers. Each output is then fed to its respective polarity switch. The D.P.D.T. switch determines to which side of the differential amplifier, V_4 , the particular signal is sent. Since more than one signal may require the same polarity, the signals are combined through "L" type resistance attenuators to prevent interaction between signals. The D.P.D.T. switches are so arranged that the other signals are unaffected when a polarity switch is reversed. The differential amplifier, V_4 , is a cathode coupled circuit having the advantage of two identical grids which produce opposing effects in the output. The output is then matched to the line by the cathode follower, V_6 . In this way we can transmit (1) the

original picture signal with either polarity and any amplitude, (2) the picture signal delayed by one or two Nyquist intervals with either polarity and any amplitude or (3) any linear combination of (1) or (2).

Fig. 12 is a block diagram of the experimental set-up used to investigate prediction methods that involve previous value and previous line samples such as "planar," etc. The input is fed into a manually variable delay having 0.1 Nyquist interval steps. This delay line acts like a vernier for the 63.5 microsecond line delay. Effectively, it enables the previous line samples to be positioned directly above the previous value samples.

The 63.5 microsecond delay is a so-called "acoustic" or "ultrasonic" delay line and was developed by Mr. H. J. McSkimin. The associated circuitry was developed by Mr. A. L. Hopper.* Storage is accomplished by a fused silica bar with quartz transducers operating at a carrier

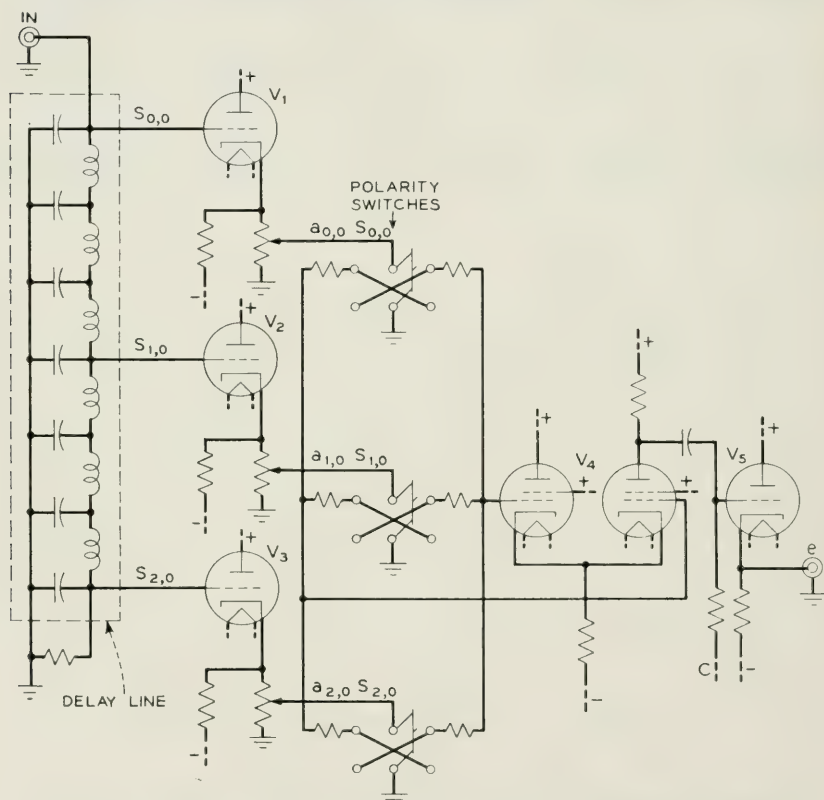


Fig. 11—Simplified schematic for "previous value" and "slope" prediction.

* A. L. Hopper, "Storing Video Information," *Electronics*, **24**, pp. 122-125, June, 1951.

BLOCK DIAGRAM OF EXPERIMENTAL SETUP

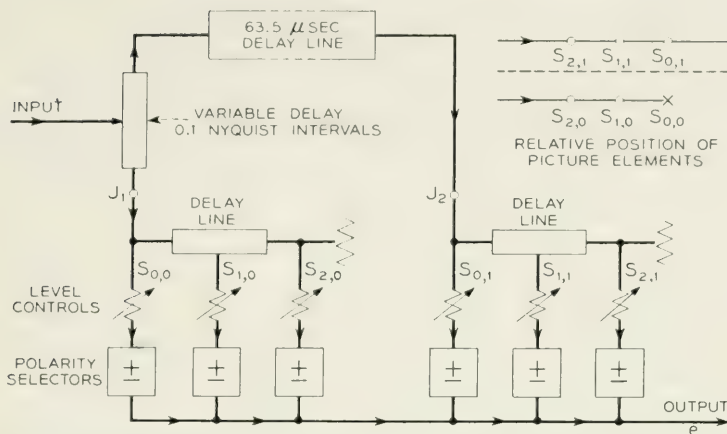


Fig. 12—Block diagram of experimental apparatus used to investigate methods of prediction involving combinations of previous signal values along a line with those on the line directly above.

frequency of 54.0 mc. The over-all video bandwidth is flat (± 0.1 db) to 5.0 mc. Nonlinear distortion is approximately one percent when the peak-to-peak signal to r.m.s. noise is 58 db. To give an idea of its complexity, two such units with their associated power supplies require a seven-foot relay rack for housing. The signal at J_1 represents the input picture signal, even though it may be delayed by a small fraction of a Nyquist interval from the actual input signal. The signal at J_2 is the same signal as found at J_1 but delayed by one line time. Each of these signals is fed into terminated delay lines to enable additional signal samples to be obtained. The geometrical location of these signal samples is illustrated in the upper right section of Fig. 12. Here, six signals are obtained instead of three as were required for "previous value" and "slope" prediction. These signals are weighted and polarized in the same manner as the three signals shown in Fig. 11. The output is the sum of these weighted signals and is given by

$$e = a_{0,0}S_{0,0} + a_{1,0}S_{1,0} + a_{2,0}S_{2,0} + a_{0,1}S_{0,1} + a_{1,1}S_{1,1} + a_{2,1}S_{2,1}.$$

The coefficients may assume positive or negative values.

MEASUREMENTS

It is obvious that if we are able to predict the value of most signal samples closely (which we will be able to do if there is a large amount of correlation in the picture), then the average amplitude of our mistakes

will be much less than the average amplitude of the original signal. Thus, by using the decorrelator alone, we can send a message over a channel with the same bandwidth as before but with less average power. At first, this might sound like a worthwhile saving; however this lower average power is accompanied by an even higher peak amplitude which makes any direct saving less attractive. Furthermore, the low frequency attenuation of the decorrelator makes the signal vulnerable to low frequency disturbances, since the correlator must restore (emphasize) these low frequency components.

A proper but *not entirely adequate* method of evaluating the effectiveness of a predictor is by measuring the ratio of signal power to error power. This is called "Power Reduction" and is generally expressed in db. Power reduction simply provides a scale by which we can weigh a linear predictor's capabilities. The "not entirely adequate" refers to the fact that minimum error power may not provide simultaneously the lowest amount of redundancy for that given type of prediction.

As an example, Fig. 13 shows the power reduction for the relative weighting of the previous horizontal signal sample as compared to the present value of the signal, for three pictures-later to be described as Scene A, B and C. The top-most curve is for Scene B, which is a simple, soft picture that contains very little detail. For this picture, the minimum error power coincides (within measurable limits) with the minimum redundancy. For Scene A and particularly Scene C, minimum error power is considerably different than that for minimum redundancy. This difference between minimum error power and minimum redundancy also applies to decorrelators using other types of predictors as well. Minimum redundancy may also be a misleading criterion of a predictor's performance, since the value of the prediction must depend on the particular type of encoder used, and some types of encoding will require certain types of redundant information to be retained.

The following pictures are representations of the error signal as photographed from a 10-inch laboratory monitor. The signals were band limited to 4.3 mc. Fig. 14 represents the "original" for three scenes called A, B and C. These pictures represent, to a first approximation, the gamut of pictures normally expected to be transmitted. They are by no means the best or the worst pictures than can be imagined; however any system should be able to reproduce these pictures without appreciable distortion. For example, Scene C should be capable of being sent continuously without the elastic delay running out, etc.

Fig. 15 shows how the error signal appears for "previous value" prediction. "Previous value" prediction is excellent for flat white or dark

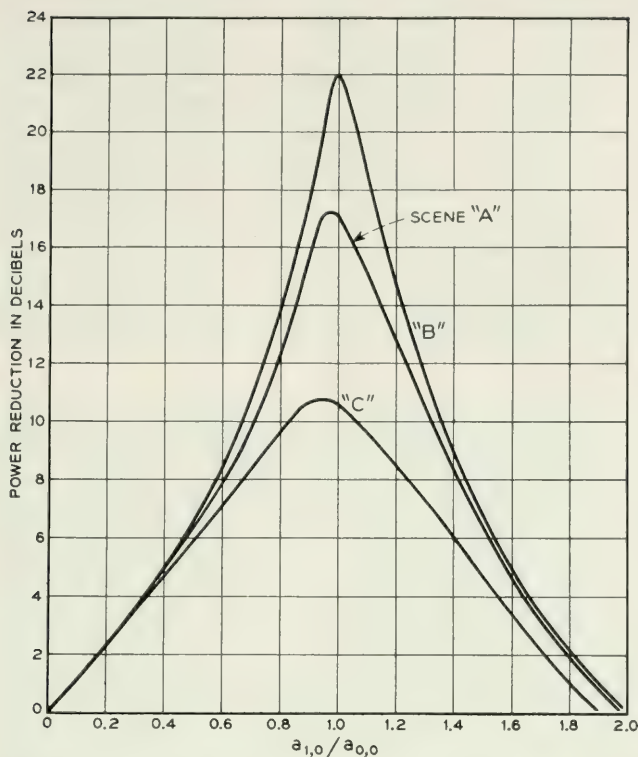


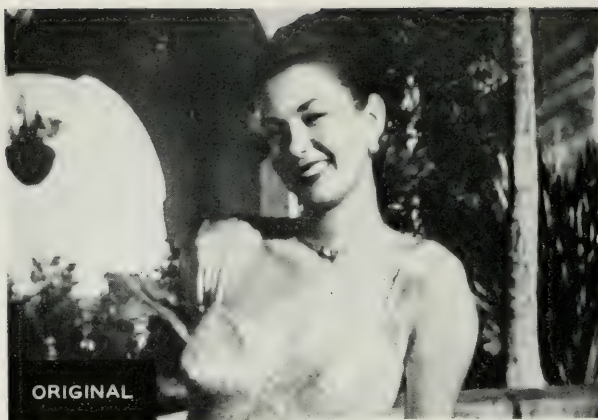
Fig. 13—Power reduction for various weighting coefficients for the previous horizontal sample.

areas as can be observed in the background of Scene A. Where the separation of a white to black area is made, the error signal is large. It is this large error signal that informs the receiver of this change in brightness, and until another change occurs, the error output is again zero. This type of performance produced the flat grey appearance of the background. In this way, the picture represents only changes in brightness—a first difference type of picture.

It may be noted that horizontal contour lines have vanished leaving only vertical contours which pertain to the brightness changes that have occurred. This effect is especially evident in Scene C. The power reductions given in the lower left hand corner of these pictures are consistent with their complexity.

Fig. 16 shows the error signal appearance for "slope" prediction. When compared to the error signal for "previous value" prediction a finer vertical granularity is observed, and this is attributed to sudden

SCENE
A



SCENE
B

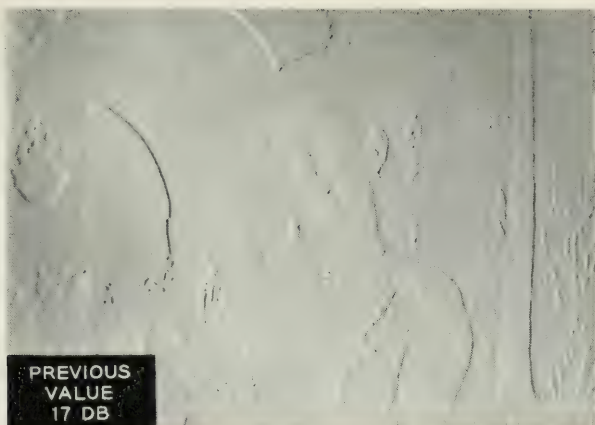


SCENE
C

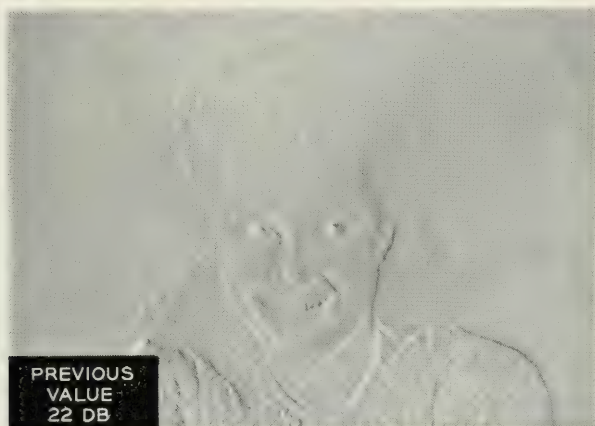


Fig. 14—Three pictures as photographed from the face of a kinescope. Scene "A" is a picture of average complexity. Scene "B" is a simple, rather soft picture. Scene "C" is a complex, highly detailed picture. Roughly, these pictures represent the gamut of pictures normally expected to be transmitted.

SCENE
A



SCENE
B

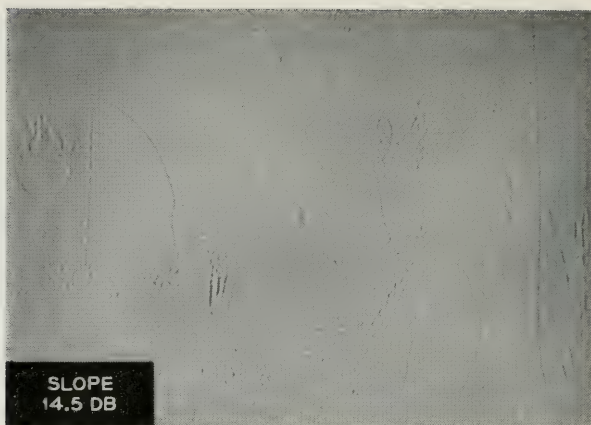


SCENE
C



Fig. 15—Three pictures showing the appearance of the error signal when using “previous value” prediction. Note the absence of horizontal contours.

SCENE
A



SCENE
B



SCENE
C

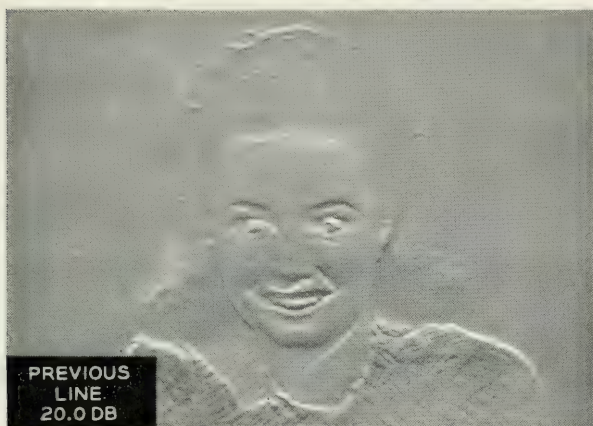


Fig. 16—Three pictures showing the appearance of the error signal when using 'slope' prediction.

SCENE
A



SCENE
B



SCENE
C

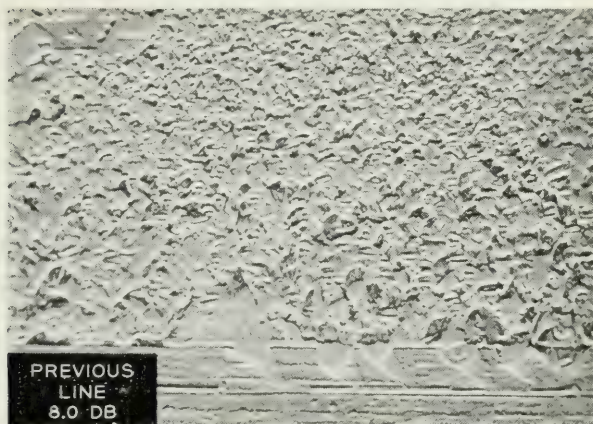
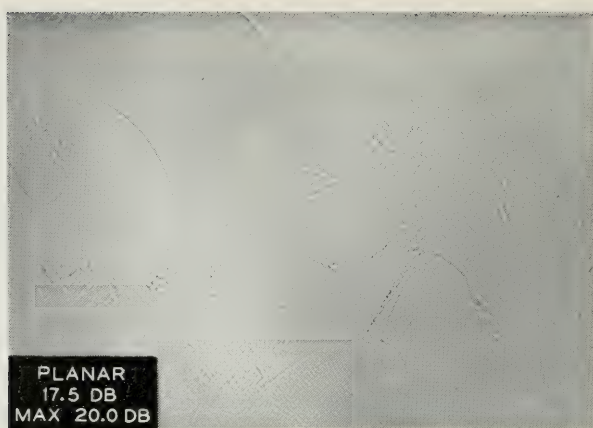
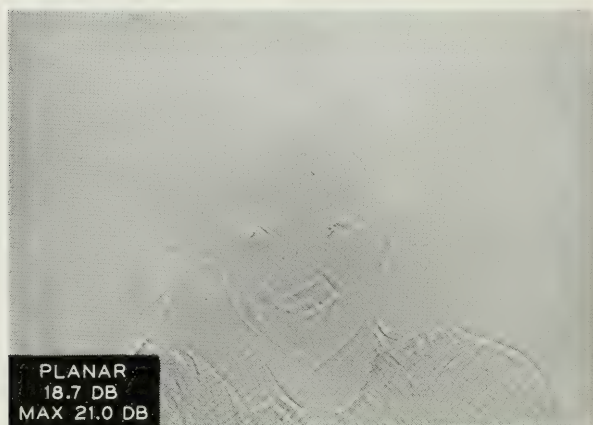


Fig. 17—Three pictures showing the appearance of the error signal when using “previous line” prediction. Note the absence of vertical contours.

SCENE
A



SCENE
B



SCENE
C

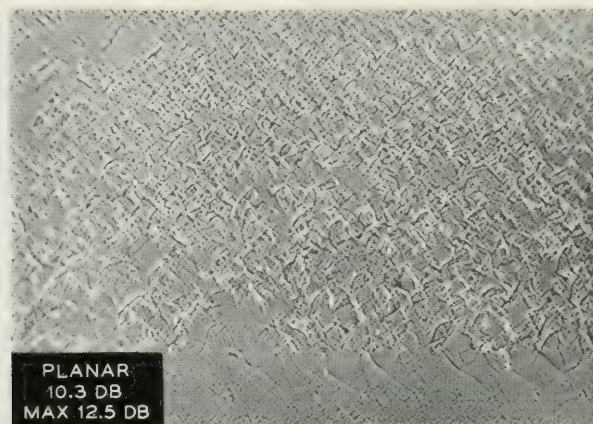


Fig. 18—Three pictures showing the appearance of the error signal when using “planar” prediction. Note the absence of horizontal and vertical contours.

changes in brightness. In the case of "previous value" prediction a sudden change in brightness produces only one error, where for "slope," two errors result. This, to some extent, accounts for the lesser amount of power reduction measured for these scenes.

Fig. 17 shows the appearance of the error signal for "previous line" prediction in the three scenes. Where vertical contour lines are predominately left after "previous value" prediction shown in Fig. 15, horizontal contours, are more prevalent now. It can be noted that the power reduction for "previous line" prediction is less than that for "previous value" prediction. This is due principally to the increased distance of the previous line sample from $S_{0,0}$. If the closest horizontal sample was taken at the same distance from the present value of the signal as the previous line sample, then the power reduction using these signal values individually for prediction would be essentially the same for most pictures.

Fig. 18 shows the error signal appearance for "planar" prediction. Here, vertical as well as horizontal contours are deleted. In Scene A the tree trunk has almost completely vanished. In Scene B the picture has an extremely flat appearance. Scene C exhibits the lack of horizontal and vertical contours best, since only sloping contours are left. The power reduction figures at the lower left hand corner also show values for minimum error power. For most pictures, the error power can be reduced by a factor of one-half again over the planar coefficients by modifying the weighting coefficients. The coefficients for this modified planar case are given by

$$S_p = \frac{2}{3}S_{1,0} + \frac{2}{3}S_{0,1} - \frac{1}{3}S_{1,1}$$

These coefficients generally produce an error signal with less power than the coefficients used for "planar" prediction.

While all pictures contain redundancy, the error signals from these simple linear predictors shown in Figs. 15, 16, 17 and 18 can visually be noted still to contain large amounts of redundancy. The contours of the models and of the various objects are readily identifiable. Were all redundancy removed, the picture would be completely chaotic and would appear very much like random noise, although greater efficiency in transmission would be achieved. For richer rewards, more sophisticated methods of prediction will be required.

ACKNOWLEDGMENT

The author wishes to acknowledge with grateful appreciation the invaluable guidance of Dr. B. M. Oliver. It was principally through his efforts that this study was made possible.

Generalized Telegraphist's Equations for Waveguides

By S. A. SCHELKUNOFF

(Manuscript received April 30, 1952)

In this paper Maxwell's partial differential equations and the boundary conditions for waveguides filled with a heterogeneous and non-isotropic medium are converted into an infinite system of ordinary differential equations. This system represents a generalization of "telegraphist's equations" for a single mode transmission to the case of multiple mode transmission. A similar set of equations is obtained for spherical waves. Although such generalized telegraphist's equations are very complicated, it is very likely that useful results can be obtained by an appropriate modal analysis.

From a purely mathematical point of view the problem of electromagnetic wave propagation inside a metal waveguide reduces to obtaining that solution of Maxwell's equations which satisfies certain boundary conditions *along* the waveguide and certain terminal conditions at the ends of the waveguide. If the medium inside the waveguide is homogeneous and isotropic and if the cross-section of the waveguide is either rectangular or circular or elliptic, the desired solution is obtained by the method of separating the variables. The method works for some other special cross-sections. It works also if the medium inside a rectangular waveguide consists of homogeneous, isotropic strata parallel to one of its faces. Similarly, it works if the medium inside a circular waveguide consists of coaxial, homogeneous, isotropic layers. But in general if the medium is either nonhomogeneous or non-isotropic or both, the method does not work. The mathematical reason for this is that the solution is of a more complicated form than a simple production of functions, each depending on a single coordinate. Any function that one usually encounters in physical problems, and therefore a solution of Maxwell's equations, may be expanded in a series of orthogonal functions. Sets of such functions are provided by the solutions for waveguides filled with homogeneous media. Such functions already satisfy the proper boundary conditions and the problem is to obtain series which also satisfy

Maxwell's equations. From the physical point of view this method represents a conversion of Maxwell's equations into generalized "telegraphist's equations."

Thus it is already known that Maxwell's partial differential equations and the boundary conditions along a waveguide are convertible into a set of independent ordinary differential equations, each resembling telegraphist's equations for electric transmission lines.¹ Each equation describes a "mode of propagation" in terms of concepts well known in electric circuit theory. A waveguide can be considered as an infinite system of transmission lines. If the medium inside the waveguide is homogeneous and isotropic and if the surface impedance of the boundary is zero, the method of separating the variables enables us to obtain a set of "normal", that is, *uncoupled* modes of propagation. Any irregularity or "discontinuity" in the waveguide provides a coupling between some, or all, modes of propagation. The irregularity may be in a boundary of the waveguide or in the dielectric within it. A heterogeneous dielectric may be considered as a homogeneous dielectric with distributed irregularities.² Similarly a heterogeneous non-isotropic dielectric may be considered as a homogeneous isotropic dielectric with distributed irregularities. Such irregularities provide a distributed coupling between the various modes appropriate to homogeneous isotropic waveguides. Our problem is to calculate the coupling coefficients. The generalized telegraphist's equations, obtained in this manner, are very complicated in that they represent an infinite number of coupled transmission modes. They are useful, however, in suggesting a physical picture of wave propagation under complicated conditions, and can be used in approximate analysis when we can ignore all but the most tightly coupled modes. For example, this picture was successfully employed by Albersheim³ in studying the effect of gentle bending of a waveguide on propagation of circular electric waves. In this case it was important to consider the coupling between only two modes, TE_{01} and TM_{11} , which have the same cutoff frequency in a straight waveguide. More recently, Stevenson obtained exact equations for waves in horns of arbitrary shape.⁴ His equations express the propagation of the axial components of E and H . The various modes are coupled through the boundary of the horn. In

¹ S. A. Schelkunoff, "Transmission Theory of Plane Electromagnetic Waves," *Proc. Inst. Radio Engrs.*, Nov. 1937, pp. 1457-1492.

² S. A. Schelkunoff, "Electromagnetic Waves," D. van Nostrand Co., (1943), pp. 92-93.

³ W. J. Albersheim, "Propagation of TE_{01} Waves in Curved Waveguides," *Bell System Tech. J.*, Jan. 1949, pp. 1-32.

⁴ A. F. Stevenson, "General Theory of Electromagnetic Horns," *J. Appl. Phys.*, Dec. 1951, pp. 1447-1460.

the present paper we shall consider waveguides of *constant* cross-section and *conical* horns of arbitrary shape filled with a heterogeneous and non-isotropic dielectric and derive the equations for propagation of the generalized voltages and currents representing the *transverse* field components. The various modes are coupled through the medium. It is very likely that our equations can be generalized to include the coupling through the boundary.

To understand the mechanism of coupling between the various modes through the medium consider Maxwell's equations

$$\text{curl } E = -j\omega B, \quad \text{curl } H = {}^c J + j\omega D, \quad (1)$$

where ${}^c J$ is the density of conduction current while the other letter symbols have the usual meanings. In the most general linear case the components of B and D are linear functions of the components of H and E respectively, with the coefficients depending on the coordinates. These equations can always be rewritten as follows

$$\text{curl } E = -j\omega\mu H - M, \quad \text{curl } H = j\omega\epsilon E + J, \quad (2)$$

where M and J are the densities of magnetic and electric polarization currents.⁵

$$M = j\omega(B - \mu H), \quad J = {}^c J + j\omega(D - \epsilon E), \quad (3)$$

and μ , ϵ are constants (not necessarily those of vacuum). If M and J were given, they would act as sources exciting various modes of propagation in a homogeneous, isotropic waveguide. If M and J are functions of H and E , they can still be considered as the sources, acting on power borrowed from the wave, of the various modes. Thus M and J will provide the coupling between the modes existing in a homogeneous, isotropic waveguide.

Thus in order to derive the generalized telegraphist's equations we shall first consider the various modes of propagation in a homogeneous isotropic wave guide. Each mode is described by a transverse field distribution pattern⁶ $T(u, v)$, where u and v are orthogonal coordinates of a point in a typical cross-section. This function is a solution of the following two-dimensional partial differential equation

$$\Delta T = \frac{1}{e_1 e_2} \left[\frac{\partial}{\partial u} \left(\frac{e_2}{e_1} \frac{\partial T}{\partial u} \right) + \frac{\partial}{\partial v} \left(\frac{e_1}{e_2} \frac{\partial T}{\partial v} \right) \right] = -\chi^2 T, \quad (4)$$

⁵ See Reference 2.

⁶ S. A. Schelkunoff, "Electromagnetic Waves," D. van Nostrand Co. (1943), Chapter 10.

where χ is the separation constant and e_1, e_2 are the scale factors in the expression for the elementary distance

$$ds^2 = e_1^2 du^2 + e_2^2 dv^2. \quad (5)$$

In the case of TM waves the T -function must vanish on the boundary of zero impedance. This boundary condition restricts χ to a doubly infinite set of values χ_{mn} with the corresponding functions T_{mn} . In the case of TE waves the normal derivative of the T -function must vanish on the boundary of zero impedance. Since we have to consider both types of waves simultaneously, we shall distinguish between them by enclosing the *subscripts in parentheses* for TM waves and *in brackets* for TE waves. The double subscript designation of various modes has been standardized only for rectangular and circular waveguides. For waveguides of other shapes the standard is to use a single subscript by arranging the modes in the order of their cutoff frequencies. For convenience, we shall use this convention in the general case and denote TM modes by $T_{(n)}(u, v)$, and TE modes by $T_{[n]}(u, v)$. The corresponding cutoff constants will be $\chi_{(n)}$ and $\chi_{[n]}$. In what follows it is understood that whenever the T -functions should be designated by double subscripts, our single letter subscripts should be considered as symbols for ordered double subscripts.

The transverse field components may be derived from the potential and stream functions,⁷ V and Π for TM waves and U and Ψ for TE waves. Thus

$$E_t = -\text{grad } V - \text{flux } \Psi, \quad H_t = \text{flux } \Pi - \text{grad } U, \quad (6)$$

where the components of the gradient and flux of a scalar function W are

$$\begin{aligned} \text{grad}_u W &= \frac{\partial W}{e_1 \partial u}, & \text{grad}_v W &= \frac{\partial W}{e_2 \partial v}, \\ \text{flux}_u W &= \frac{\partial W}{e_2 \partial v}, & \text{flux}_v W &= -\frac{\partial W}{e_1 \partial u}. \end{aligned} \quad (7)$$

The T -functions corresponding to the various modes of the same variety are orthogonal; that is, the following integrals over the cross-section vanish,

$$\iint T_{(n)} T_{(m)} dS = 0, \quad \iint T_{[n]} T_{[m]} dS = 0, \quad \text{if } m \neq n. \quad (8)$$

It should be stressed that $T_{(n)}$ and $T_{[m]}$ are not, in general, orthogonal.

⁷ See Reference 6.

Similarly the gradients of the T -functions of the same variety as well as the fluxes, are orthogonal,

$$\begin{aligned} \iint (\text{grad } T_{(n)}) \cdot (\text{grad } T_{(m)}) dS &= \iint (\text{flux } T_{(n)}) \cdot (\text{flux } T_{(m)}) dS \\ &= \iint (\text{grad } T_{[n]}) \cdot (\text{grad } T_{[m]}) dS = \iint (\text{flux } T_{[n]}) \cdot (\text{flux } T_{[m]}) dS = 0, \end{aligned} \quad (9)$$

if $m \neq n$. The following gradients and fluxes of the T -functions are orthogonal for all m and n ,

$$\begin{aligned} \iint (\text{grad } T_{(n)}) \cdot (\text{flux } T_{[m]}) dS &= \iint (\text{grad } T_{[n]}) \cdot (\text{flux } T_{(m)}) dS \\ &= \iint (\text{grad } T_{(n)}) \cdot (\text{flux } T_{(m)}) dS = 0. \end{aligned} \quad (10)$$

On the other hand, $\text{grad } T_{[m]}$ and $\text{flux } T_{[n]}$ are not, in general, orthogonal.

If all modes are present, the potential and stream functions are

$$\begin{aligned} V &= -V_{(n)}(z)T_{(n)}(u, v), & \Pi &= -I_{(n)}(z)T_{(n)}(u, v), \\ \Psi &= -V_{[n]}(z)T_{[n]}(u, v), & U &= -I_{[n]}(z)T_{[n]}(u, v), \end{aligned} \quad (11)$$

where the tensor summation convention is used: whenever the same letter subscript is used in a product, it should receive all values in a given set and the resulting products should be added. The negative signs have been inserted in order to avoid a preponderance of negative signs in later equations. Substituting in (9), we have

$$\begin{aligned} E_t &= V_{(n)} \text{grad } T_{(n)} + V_{[n]} \text{flux } T_{[n]}, \\ H_t &= -I_{(n)} \text{flux } T_{(n)} + I_{[n]} \text{grad } T_{[n]}. \end{aligned} \quad (12)$$

The T -functions for the various modes are determined by equation (4) and the boundary conditions except for arbitrary factors related to the power levels of the modes. If we choose these constants in such a way that

$$\iint (\text{grad } T) \cdot (\text{grad } T) dS \equiv \chi^2 \iint T^2 dS = 1, \quad (13)$$

then the complex power carried by the wave is given by an expression similar to that in an ordinary transmission line,

$$P = \frac{1}{2} V_{(n)} I_{(n)}^* + \frac{1}{2} V_{[n]} I_{[n]}^*. \quad (14)$$

Hence, the V 's and I 's correspond to the voltages and currents in coupled transmission lines.

In an expanded form equations (12) are

$$\begin{aligned} E_u &= V_{(n)} \frac{\partial T_{(n)}}{e_1 \partial u} + V_{[n]} \frac{\partial T_{[n]}}{e_2 \partial v}, & E_v &= V_{(n)} \frac{\partial T_{(n)}}{e_2 \partial v} - V_{[n]} \frac{\partial T_{[n]}}{e_1 \partial u}, \\ H_v &= I_{(n)} \frac{\partial T_{(n)}}{e_1 \partial u} + I_{[n]} \frac{\partial T_{[n]}}{e_2 \partial v}, & H_u &= -I_{(n)} \frac{\partial T_{(n)}}{e_2 \partial v} + I_{[n]} \frac{\partial T_{[n]}}{e_1 \partial u}. \end{aligned} \quad (15)$$

To these we add the following expansions for the longitudinal components of E and H

$$E_z = \chi_{(n)} V_{z,(n)}(z) T_{(n)}(u, v), \quad H_z = \chi_{[n]} I_{z,[n]}(z) T_{[n]}(u, v). \quad (16)$$

Equations of this form satisfy automatically the boundary conditions on E_z and H_z . The multipliers χ_n have been inserted arbitrarily in order to make the physical dimensions of the second factors to correspond to those of voltage and current.

Let us now write Maxwell's equations in an expanded form

$$\begin{aligned} \frac{\partial E_z}{e_2 \partial v} - \frac{\partial E_v}{\partial z} &= -j\omega B_u, & \frac{\partial H_z}{e_2 \partial v} - \frac{\partial H_v}{\partial z} &= j\omega D_u, \\ \frac{\partial E_u}{\partial z} - \frac{\partial E_z}{e_1 \partial u} &= -j\omega B_v, & \frac{\partial H_u}{\partial z} - \frac{\partial H_z}{e_1 \partial u} &= j\omega D_v, \\ \frac{\partial(e_2 E_v)}{\partial u} - \frac{\partial(e_1 E_u)}{\partial v} &= -j\omega e_1 e_2 B_z, & \frac{\partial(e_2 H_v)}{\partial u} - \frac{\partial(e_1 H_u)}{\partial v} &= j\omega e_1 e_2 D_z. \end{aligned} \quad (17)$$

Substituting from (15) and (16) in the left column of (17), we find

$$\chi_{(n)} V_{z,(n)} \frac{\partial T_{(n)}}{e_2 \partial v} - \frac{dV_{(n)}}{dz} \frac{\partial T_{(n)}}{e_2 \partial v} + \frac{dV_{[n]}}{dz} \frac{\partial T_{[n]}}{e_1 \partial u} = -j\omega B_u, \quad (18)$$

$$-\chi_{(n)} V_{z,(n)} \frac{\partial T_{(n)}}{e_1 \partial u} + \frac{dV_{(n)}}{dz} \frac{\partial T_{(n)}}{e_1 \partial u} + \frac{dV_{[n]}}{dz} \frac{\partial T_{[n]}}{e_2 \partial v} = -j\omega B_v, \quad (19)$$

$$\begin{aligned} V_{(n)} \frac{\partial^2 T_{(n)}}{\partial u \partial v} - V_{[n]} \frac{\partial}{\partial u} \left(\frac{e_2}{e_1} \frac{\partial T_{[n]}}{\partial u} \right) - V_{(n)} \frac{\partial^2 T_{(n)}}{\partial v \partial u} - V_{[n]} \frac{\partial}{\partial v} \left(\frac{e_1}{e_2} \frac{\partial T_{[n]}}{\partial v} \right) \\ = -j\omega e_1 e_2 B_z. \end{aligned} \quad (20)$$

In view of (4) the last equation reduces to

$$\chi_{[n]}^2 V_{[n]} T_{[n]} = -j\omega B_z. \quad (21)$$

Multiplying (18) by $[-\partial T_{(m)}/e_2 \partial v] dS$, (19) by $[\partial T_{(m)}/e_1 \partial u] dS$, adding, and integrating over the cross-section, we obtain

$$-\chi_{(m)} V_{z,(m)} + \frac{dV_{(m)}}{dz} = j\omega \iint \left(B_u \frac{\partial T_{(m)}}{e_2 \partial v} - B_v \frac{\partial T_{(m)}}{e_1 \partial u} \right) dS. \quad (22)$$

In the first term the summation convention should be ignored. Multiplying (18) by $[\partial T_{[m]}/e_1 \partial u] dS$, (19) by $[\partial T_{[m]}/e_2 \partial v] dS$, adding, and integrating we find

$$\frac{\partial V_{[m]}}{dz} = -j\omega \iint \left(B_u \frac{\partial T_{[m]}}{e_1 \partial u} + B_v \frac{\partial T_{[m]}}{e_2 \partial v} \right) dS. \quad (23)$$

Multiplying (21) by $T_{[m]} dS$ and integrating, we have

$$V_{[m]} = -j\omega \iint B_z T_{[m]} dS. \quad (24)$$

Subjecting the right column of (17) to a similar treatment, we obtain three additional equations. Summarizing, we have

$$\frac{\partial V_{(m)}}{dz} = j\omega \iint \left(B_u \frac{\partial T_{(m)}}{e_2 \partial v} - B_v \frac{\partial T_{(m)}}{e_1 \partial u} \right) dS + \chi_{(m)} V_{z,(m)}, \quad (25)$$

$$\frac{\partial I_{(m)}}{dz} = -j\omega \iint \left(D_u \frac{\partial T_{(m)}}{e_1 \partial u} + D_v \frac{\partial T_{(m)}}{e_2 \partial v} \right) dS, \quad (26)$$

$$\frac{dV_{[m]}}{dz} = -j\omega \iint \left(B_u \frac{\partial T_{[m]}}{e_1 \partial u} + B_v \frac{\partial T_{[m]}}{e_2 \partial v} \right) dS, \quad (27)$$

$$\frac{dI_{[m]}}{dz} = j\omega \iint \left(-D_u \frac{\partial T_{[m]}}{e_2 \partial v} + D_v \frac{\partial T_{[m]}}{e_1 \partial u} \right) dS + \chi_{[m]} I_{z,[m]}, \quad (28)$$

$$V_{[m]} = -j\omega \iint B_z T_{[m]} dS, \quad I_{(m)} = -j\omega \iint D_z T_{(m)} dS. \quad (29)$$

In the last terms of equations (25) and (28) the summation convention should be ignored.

If the components of B and D are linear functions of the components of H and E respectively, then with the aid of (15) and (16) they can be expressed as linear functions of $V_{(n)}$, $V_{[n]}$, $I_{(n)}$, $I_{[n]}$, $V_{z,(n)}$, $I_{z,[n]}$. Solving (29) for $V_{z,(n)}$ and $I_{z,[n]}$ and making the appropriate substitutions in (25), (26), (27), (28), we obtain the generalized telegraphist's

equations in the following form

$$\begin{aligned}
 \frac{dV_{(m)}}{dz} &= -Z_{(m)(n)}I_{(n)} - Z_{(m)[n]}I_{[n]} - {}^V T_{(m)(n)}V_{(n)} - {}^V T_{(m)[n]}V_{[n]}, \\
 \frac{dI_{(m)}}{dz} &= -Y_{(m)(n)}V_{(n)} - Y_{(m)[n]}V_{[n]} - {}^I T_{(m)(n)}I_{(n)} - {}^I T_{(m)[n]}I_{[n]}, \\
 \frac{dV_{[m]}}{dz} &= -Z_{[m][n]}I_{[n]} - Z_{[m](n)}I_{(n)} - {}^V T_{[m][n]}V_{[n]} - {}^V T_{[m](n)}V_{(n)}, \\
 \frac{dI_{[m]}}{dz} &= -Y_{[m][n]}V_{[n]} - Y_{[m](n)}V_{(n)} - {}^I T_{[m][n]}I_{[n]} - {}^I T_{[m](n)}I_{(n)}.
 \end{aligned} \tag{30}$$

The transfer impedances Z , the transfer admittances Y , the voltage transfer coefficients ${}^V T$, and the current transfer coefficients ${}^I T$ between various modes are in general functions of z . They are constants if the properties of the waveguide are independent of the distance along it; in this case the problem of solving the generalized telegraphist's equations reduces to solving an infinite system of linear algebraic equations and the corresponding characteristic equation.

Similar equations may be derived for spherical waves either in an unlimited medium or in a medium bounded by a perfectly conducting conical surface of arbitrary cross-section. If the latter is circular and if the flare angle is 180° , we have a plane boundary. Hence, the case of spherical waves in a non-homogeneous medium is included. In the spherical case we shall use the general orthogonal system of coordinates (r, u, v) where r is the distance from the center and (u, v) are orthogonal angular coordinates. In this system the elements of length ds and area dS are given by

$$ds^2 = dr^2 + r^2(e_1^2 du^2 + e_2^2 dv^2), \quad dS = r^2 d\Omega, \quad d\Omega = e_1 e_2 du dv. \tag{31}$$

The transverse field components may be expressed in a form similar to that for waveguides

$$rE_t = -\text{grad } V - \text{flux } \Pi, \quad rH_t = \text{flux } \Pi - \text{grad } U, \tag{32}$$

where grad and flux of a typical scalar function are defined by equations (10). Instead of (11) we have

$$\begin{aligned}
 V &= -V_{(n)}(r)T_{(n)}(u, v), & \Pi &= -I_{(n)}(r)T_{(n)}(u, v), \\
 \Psi &= -V_{[n]}(r)T_{[n]}(u, v), & U &= -I_{[n]}(r)T_{[n]}(u, v),
 \end{aligned} \tag{33}$$

where the T -functions satisfy equation (4) and appropriate boundary conditions. These functions, their gradients and fluxes are orthogonal.

They are assumed to be normalized as follows

$$\iint (\text{grad } T) \cdot (\text{grad } T) d\Omega = \chi^2 \iint T^2 d\Omega = 1, \quad (34)$$

where $d\Omega$ is an elementary solid angle. Hence, equation (14) will again represent the complex power flow in the direction of propagation.

The various field components may then be expressed as follows

$$\begin{aligned} rE_u &= V_{(n)} \frac{\partial T_{(n)}}{e_1 \partial u} + V_{[n]} \frac{\partial T_{[n]}}{e_2 \partial v}, & rE_v &= V_{(n)} \frac{\partial T_{(n)}}{e_2 \partial v} - V_{[n]} \frac{\partial T_{[n]}}{e_1 \partial u}, \\ rH_v &= I_{(n)} \frac{\partial T_{(n)}}{e_1 \partial u} + I_{[n]} \frac{\partial T_{[n]}}{e_2 \partial v}, & rH_u &= -I_{(n)} \frac{\partial T_{(n)}}{e_2 \partial v} + I_{[n]} \frac{\partial T_{[n]}}{e_1 \partial u}, \\ r^2 E_r &= \chi_{(n)} V_{r,(n)} T_{(n)}, & r^2 H_r &= \chi_{[n]} I_{r,[n]} T_{[n]}. \end{aligned} \quad (35)$$

It should be noted that the physical dimensions of $V_{r,(n)}$ and $I_{r,[n]}$ are not those of voltage and current. Substituting in Maxwell's equations and using transformations similar to those in the case of plane waves, we find

$$\begin{aligned} \frac{dV_{(m)}}{dr} &= j\omega \iint \left(rB_u \frac{\partial T_{(m)}}{e_2 \partial v} - rB_v \frac{\partial T_{(m)}}{e_1 \partial u} \right) d\Omega + \chi_{(m)} r^{-2} V_{r,(m)}, \\ \frac{dI_{(m)}}{dr} &= -j\omega \iint \left(rD_u \frac{\partial T_{(m)}}{e_1 \partial u} + rD_v \frac{\partial T_{(m)}}{e_2 \partial v} \right) d\Omega, \\ \frac{dV_{[m]}}{dr} &= -j\omega \iint \left(rB_u \frac{\partial T_{[m]}}{e_1 \partial u} + rB_v \frac{\partial T_{[m]}}{e_2 \partial v} \right) d\Omega, \\ \frac{dI_{[m]}}{dr} &= j\omega \iint \left(-rD_u \frac{\partial T_{[m]}}{e_2 \partial v} + rD_v \frac{\partial T_{[m]}}{e_1 \partial u} \right) d\Omega + \chi_{[m]} r^{-2} I_{r,[m]}, \\ V_{[m]} &= -j\omega \iint (r^2 B_r) T_{[m]} d\Omega, & I_{(m)} &= -j\omega \iint (r^2 D_r) T_{(m)} d\Omega. \end{aligned} \quad (36)$$

Returning to the plane wave case and assuming the following general linear relations

$$\begin{aligned} B_u &= \mu_{uu} H_u + \mu_{uv} H_v + \mu_{uz} H_z, & D_u &= \epsilon_{uu} E_u + \epsilon_{uv} E_v + \epsilon_{uz} E_z, \\ B_v &= \mu_{vu} H_u + \mu_{vv} H_v + \mu_{vz} H_z, & D_v &= \epsilon_{vu} E_u + \epsilon_{vv} E_v + \epsilon_{vz} E_z, \\ B_z &= \mu_{zu} H_u + \mu_{zv} H_v + \mu_{zz} H_z, & D_z &= \epsilon_{zu} E_u + \epsilon_{zv} E_v + \epsilon_{zz} E_z, \end{aligned} \quad (37)$$

we find

$$\begin{aligned}
 B_u &= I_{(n)} \left[-\mu_{uu} \frac{\partial T_{(n)}}{e_2 \partial v} + \mu_{uv} \frac{\partial T_{(n)}}{e_1 \partial u} \right] + I_{[n]} \left[\mu_{uu} \frac{\partial T_{[n]}}{e_1 \partial u} + \mu_{uv} \frac{\partial T_{[n]}}{e_2 \partial v} \right] \\
 &\quad + I_{z,[n]} \mu_{uz} \chi_{[n]} T_{[n]}, \\
 B_v &= I_{(n)} \left[-\mu_{vu} \frac{\partial T_{(n)}}{e_2 \partial v} + \mu_{vv} \frac{\partial T_{(n)}}{e_1 \partial u} \right] + I_{[n]} \left[\mu_{vu} \frac{\partial T_{[n]}}{e_1 \partial u} + \mu_{vv} \frac{\partial T_{[n]}}{e_2 \partial v} \right] \\
 &\quad + I_{z,[n]} \mu_{vz} \chi_{[n]} T_{[n]}, \\
 B_z &= I_{(n)} \left[-\mu_{zu} \frac{\partial T_{(n)}}{e_2 \partial v} + \mu_{zv} \frac{\partial T_{(n)}}{e_1 \partial u} \right] + I_{[n]} \left[\mu_{zu} \frac{\partial T_{[n]}}{e_1 \partial u} + \mu_{zv} \frac{\partial T_{[n]}}{e_2 \partial v} \right] \\
 &\quad + I_{z,[n]} \mu_{zz} \chi_{[n]} T_{[n]}, \tag{38} \\
 D_u &= V_{(n)} \left[\epsilon_{uu} \frac{\partial T_{(n)}}{e_1 \partial u} + \epsilon_{uv} \frac{\partial T_{(n)}}{e_2 \partial v} \right] + V_{[n]} \left[\epsilon_{uu} \frac{\partial T_{[n]}}{e_2 \partial v} - \epsilon_{uv} \frac{\partial T_{[n]}}{e_1 \partial u} \right] \\
 &\quad + V_{z,(n)} \epsilon_{uz} \chi_{(n)} T_{(n)}, \\
 D_v &= V_{(n)} \left[\epsilon_{vu} \frac{\partial T_{(n)}}{e_1 \partial u} + \epsilon_{vv} \frac{\partial T_{(n)}}{e_2 \partial v} \right] + V_{[n]} \left[\epsilon_{vu} \frac{\partial T_{[n]}}{e_2 \partial v} - \epsilon_{vv} \frac{\partial T_{[n]}}{e_1 \partial u} \right] \\
 &\quad + V_{z,(n)} \epsilon_{vz} \chi_{(n)} T_{(n)}, \\
 D_z &= V_{(n)} \left[\epsilon_{zu} \frac{\partial T_{(n)}}{e_1 \partial u} + \epsilon_{zv} \frac{\partial T_{(n)}}{e_2 \partial v} \right] + V_{[n]} \left[\epsilon_{zu} \frac{\partial T_{[n]}}{e_2 \partial v} - \epsilon_{zv} \frac{\partial T_{[n]}}{e_1 \partial u} \right] \\
 &\quad + V_{z,(n)} \epsilon_{zz} \chi_{(n)} T_{(n)}.
 \end{aligned}$$

Substituting from equations (38) into equations (25) to (29) we obtain

$$\begin{aligned}
 \frac{dV_{(m)}}{dz} &= -j\omega I_{(n)} \iint \left[\mu_{uu} \frac{\partial T_{(n)}}{e_2 \partial v} \frac{\partial T_{(m)}}{e_2 \partial v} + \mu_{vv} \frac{\partial T_{(n)}}{e_1 \partial u} \frac{\partial T_{(m)}}{e_1 \partial u} \right. \\
 &\quad \left. - \mu_{uv} \frac{\partial T_{(n)}}{e_1 \partial u} \frac{\partial T_{(m)}}{e_2 \partial v} - \mu_{vu} \frac{\partial T_{(n)}}{e_2 \partial v} \frac{\partial T_{(m)}}{e_1 \partial u} \right] dS \\
 &\quad + j\omega I_{[n]} \iint \left[\mu_{uu} \frac{\partial T_{[n]}}{e_1 \partial u} \frac{\partial T_{(m)}}{e_2 \partial v} - \mu_{vv} \frac{\partial T_{[n]}}{e_2 \partial v} \frac{\partial T_{(m)}}{e_1 \partial u} \right. \\
 &\quad \left. + \mu_{uv} \frac{\partial T_{[n]}}{e_2 \partial v} \frac{\partial T_{(m)}}{e_2 \partial v} - \mu_{vu} \frac{\partial T_{[n]}}{e_1 \partial u} \frac{\partial T_{(m)}}{e_1 \partial u} \right] dS \\
 &\quad + j\omega I_{z,[n]} \iint \left[\mu_{uz} \frac{\partial T_{(m)}}{e_2 \partial v} - \mu_{vz} \frac{\partial T_{(m)}}{e_1 \partial u} \right] \chi_{[n]} T_{[n]} dS + \chi_{(m)} V_{z,(m)}, \tag{39}
 \end{aligned}$$

$$\begin{aligned}
\frac{dI_{(m)}}{dz} = & -j\omega V_{(n)} \iint \left[\epsilon_{uu} \frac{\partial T_{(n)}}{e_1 \partial u} \frac{\partial T_{(m)}}{e_1 \partial u} + \epsilon_{vv} \frac{\partial T_{(n)}}{e_2 \partial v} \frac{\partial T_{(m)}}{e_2 \partial v} \right. \\
& \left. + \epsilon_{uv} \frac{\partial T_{(n)}}{e_2 \partial v} \frac{\partial T_{(m)}}{e_1 \partial u} + \epsilon_{vu} \frac{\partial T_{(n)}}{e_1 \partial u} \frac{\partial T_{(m)}}{e_2 \partial v} \right] dS \\
& + j\omega V_{[n]} \iint \left[-\epsilon_{uu} \frac{\partial T_{[n]}}{e_2 \partial v} \frac{\partial T_{(m)}}{e_1 \partial u} + \epsilon_{vv} \frac{\partial T_{[n]}}{e_1 \partial u} \frac{\partial T_{(m)}}{e_2 \partial v} \right. \\
& \left. + \epsilon_{uv} \frac{\partial T_{[n]}}{e_1 \partial u} \frac{\partial T_{(m)}}{e_1 \partial u} - \epsilon_{vu} \frac{\partial T_{[n]}}{e_2 \partial v} \frac{\partial T_{(m)}}{e_2 \partial v} \right] dS
\end{aligned} \quad (40)$$

$$\begin{aligned}
& - j\omega V_{z,(n)} \iint \left[\epsilon_{uz} \frac{\partial T_{(m)}}{e_1 \partial u} + \epsilon_{vz} \frac{\partial T_{(m)}}{e_2 \partial v} \right] \chi_{(n)} T_{(n)} dS, \\
\frac{dV_{[m]}}{dz} = & j\omega I_{(n)} \iint \left[\mu_{uu} \frac{\partial T_{(n)}}{e_2 \partial v} \frac{\partial T_{[m]}}{e_1 \partial u} - \mu_{vv} \frac{\partial T_{(n)}}{e_1 \partial u} \frac{\partial T_{[m]}}{e_2 \partial v} \right. \\
& \left. - \mu_{uv} \frac{\partial T_{(n)}}{e_1 \partial u} \frac{\partial T_{[m]}}{e_1 \partial u} + \mu_{vu} \frac{\partial T_{(n)}}{e_2 \partial v} \frac{\partial T_{[m]}}{e_2 \partial v} \right] dS \\
& - j\omega I_{[n]} \iint \left[\mu_{uu} \frac{\partial T_{[n]}}{e_1 \partial u} \frac{\partial T_{[m]}}{e_1 \partial u} + \mu_{vv} \frac{\partial T_{[n]}}{e_2 \partial v} \frac{\partial T_{[m]}}{e_2 \partial v} \right. \\
& \left. + \mu_{uv} \frac{\partial T_{[n]}}{e_2 \partial v} \frac{\partial T_{[m]}}{e_1 \partial u} + \mu_{vu} \frac{\partial T_{[n]}}{e_1 \partial u} \frac{\partial T_{[m]}}{e_2 \partial v} \right] dS
\end{aligned} \quad (41)$$

$$\begin{aligned}
& - j\omega I_{z,[n]} \iint \left[\mu_{uz} \frac{\partial T_{[m]}}{e_1 \partial u} + \mu_{vz} \frac{\partial T_{[m]}}{e_2 \partial v} \right] \chi_{[n]} T_{[n]} dS, \\
\frac{dI_{[m]}}{dz} = & j\omega V_{(n)} \iint \left[-\epsilon_{uu} \frac{\partial T_{(n)}}{e_1 \partial u} \frac{\partial T_{[m]}}{e_2 \partial v} + \epsilon_{vv} \frac{\partial T_{(n)}}{e_2 \partial v} \frac{\partial T_{[m]}}{e_1 \partial u} \right. \\
& \left. - \epsilon_{uv} \frac{\partial T_{(n)}}{e_2 \partial v} \frac{\partial T_{[m]}}{e_2 \partial v} + \epsilon_{vu} \frac{\partial T_{(n)}}{e_1 \partial u} \frac{\partial T_{[m]}}{e_1 \partial u} \right] dS \\
& - j\omega V_{[n]} \iint \left[\epsilon_{uu} \frac{\partial T_{[n]}}{e_2 \partial v} \frac{\partial T_{[m]}}{e_2 \partial v} + \epsilon_{vv} \frac{\partial T_{[n]}}{e_1 \partial u} \frac{\partial T_{[m]}}{e_1 \partial u} \right. \\
& \left. - \epsilon_{uv} \frac{\partial T_{[n]}}{e_1 \partial u} \frac{\partial T_{[m]}}{e_2 \partial v} - \epsilon_{vu} \frac{\partial T_{[n]}}{e_2 \partial v} \frac{\partial T_{[m]}}{e_1 \partial u} \right] dS
\end{aligned} \quad (42)$$

$$\begin{aligned}
& + j\omega V_{z,(n)} \iint \left[-\epsilon_{uz} \frac{\partial T_{[m]}}{e_2 \partial v} + \epsilon_{vz} \frac{\partial T_{[m]}}{e_1 \partial u} \right] \chi_{(n)} T_{(n)} dS + \chi_{[m]} I_{z,[m]}, \\
I_{z,[n]} \iint j\omega \mu_{zz} \chi_{[n]} T_{[n]} T_{[m]} dS = & -V_{[m]} \\
& + I_{(p)} \iint j\omega \left[\mu_{zu} \frac{\partial T_{(p)}}{e_2 \partial v} - \mu_{zv} \frac{\partial T_{(p)}}{e_1 \partial u} \right] T_{[m]} dS \\
& - I_{[p]} \iint j\omega \left[\mu_{zu} \frac{\partial T_{[p]}}{e_1 \partial u} + \mu_{zv} \frac{\partial T_{[p]}}{e_2 \partial v} \right] T_{[m]} dS,
\end{aligned} \quad (43)$$

$$\begin{aligned}
 V_{z,(n)} \iint j\omega \epsilon_{zz} \chi_{(n)} T_{(n)} T_{(m)} dS &= -I_{(m)} \\
 &- V_{(p)} \iint j\omega \left[\epsilon_{zu} \frac{\partial T_{(p)}}{e_1 \partial u} + \epsilon_{zv} \frac{\partial T_{(p)}}{e_2 \partial v} \right] T_{(m)} dS \\
 &+ V_{[p]} \iint j\omega \left[-\epsilon_{zu} \frac{\partial T_{[p]}}{e_2 \partial v} + \epsilon_{zv} \frac{\partial T_{[p]}}{e_1 \partial u} \right] T_{(m)} dS.
 \end{aligned} \quad (44)$$

If we solve the last two equations for $I_{z,[n]}$ and $V_{z,(n)}$ and substitute in the preceding four equations, we shall obtain the telegraphist's equations in their final form (30). Thus, let

$$\begin{aligned}
 {}^z Z_{[m][n]} &= \iint j\omega \mu_{zz} \chi_{[n]} T_{[n]} T_{[m]} dS, \\
 {}^z Y_{(m)(n)} &= \iint j\omega \epsilon_{zz} \chi_{(n)} T_{(n)} T_{(m)} dS.
 \end{aligned} \quad (45)$$

From these coefficients we obtain another set

$$\begin{aligned}
 {}^z Z_{[n][m]} &= \text{normalized co-factor of } {}^z Z_{[m][n]}, \\
 {}^z Z_{(n)(m)} &= \text{normalized co-factor of } {}^z Y_{(m)(n)}.
 \end{aligned} \quad (46)$$

Then,

$$\begin{aligned}
 I_{z,[n]} &= -V_{[m]} {}^z Y_{[n][m]} \\
 &+ I_{(p)} {}^z Y_{[n][m]} \iint j\omega \left[\mu_{zu} \frac{\partial T_{(p)}}{e_2 \partial v} - \mu_{zv} \frac{\partial T_{(p)}}{e_1 \partial u} \right] T_{[m]} dS \\
 &- I_{[p]} {}^z Y_{[n][m]} \iint j\omega \left[\mu_{zu} \frac{\partial T_{[p]}}{e_1 \partial u} + \mu_{zv} \frac{\partial T_{[p]}}{e_2 \partial v} \right] T_{[m]} dS,
 \end{aligned} \quad (47)$$

$$\begin{aligned}
 V_{z,(n)} &= -I_{(m)} {}^z Z_{(n)(m)} \\
 &- V_{(p)} {}^z Z_{(n)(m)} \iint j\omega \left[\epsilon_{zu} \frac{\partial T_{(p)}}{e_1 \partial u} + \epsilon_{zv} \frac{\partial T_{(p)}}{e_2 \partial v} \right] T_{(m)} dS \\
 &+ V_{[p]} {}^z Z_{(n)(m)} \iint j\omega \left[-\epsilon_{zu} \frac{\partial T_{[p]}}{e_2 \partial v} + \epsilon_{zv} \frac{\partial T_{[p]}}{e_1 \partial u} \right] T_{(m)} dS.
 \end{aligned}$$

Before substituting in equations (39) to (42), the summation index m in (47) should be changed to avoid conflict with m in the former equations. It does not seem necessary to make these final substitutions in their most general form. The results are very complicated and in practice the various coefficients are not independent. Some coefficients may

vanish; others may be small. In isotropic media, $\mu_{uu} = \mu_{vv} = \mu_{zz} = \mu$, $\epsilon_{uu} = \epsilon_{vv} = \epsilon_{zz} = \epsilon$ and the mutual coefficients vanish. In gyromagnetic media subjected to a strong magnetic field in the z -direction, the permeability coefficients of superposed ac fields are⁸

$$\mu_{uu} = \mu_{vv} = \mu, \quad \mu_{vu} = -\mu_{uv}, \quad \mu_{zu} = \mu_{zv} = \mu_{uz} = \mu_{vz} = 0. \quad (48)$$

If the entire waveguide is filled with such a medium, assumed to be homogeneous, equations (43) and (44) become

$$\begin{aligned} I_{z,[n]} j\omega \mu_{zz} \chi_{[n]} \iint T_{[n]} T_{[m]} dS &= -V_{[m]}, \\ V_{z,(n)} j\omega \epsilon \chi_{(n)} \iint T_{(n)} T_{(m)} dS &= -I_{(m)}. \end{aligned} \quad (49)$$

In view of the orthogonality of the T -functions and the normalization conditions (13), we have

$$I_{z,[m]} = -\frac{\chi_{[m]}}{j\omega \mu_{zz}} V_{[m]}, \quad V_{z,(m)} = -\frac{\chi_{(m)}}{j\omega \epsilon} I_{(m)}, \quad (50)$$

where the summation convention is waived. In this case all the transfer coefficients in equations (30) vanish,

$$\begin{aligned} {}^v T_{(m)(n)} = {}^v T_{(m)[n]} = {}^v T_{[m][n]} = {}^v T_{[m](n)} = {}^I T_{(m)(n)} = {}^I T_{(m)[n]} \\ = {}^I T_{[m](n)} = {}^I T_{[m][n]} = 0. \end{aligned} \quad (51)$$

The transfer impedances and admittances are

$$\begin{aligned} Z_{(m)(n)} &= 0, \quad \text{if } n \neq m, \\ &= j\omega \mu + \frac{\chi_{(m)}^2}{j\omega \epsilon}, \quad \text{if } n = m; \\ Z_{(m)[n]} &= -j\omega \mu_{uv} \iint \left[\frac{\partial T_{[n]}}{\partial e_1} \frac{\partial T_{(m)}}{\partial u} + \frac{\partial T_{[n]}}{\partial e_2} \frac{\partial T_{(m)}}{\partial v} \right] e_1 e_2 du dv; \\ Y_{(m)(n)} &= 0, \quad \text{if } n \neq m, \\ &= j\omega \epsilon, \quad \text{if } n = m; \\ Y_{(m)[n]} &= 0, \quad \text{all } m, n; \\ Z_{[m][n]} &= j\omega \mu_{uv} \iint \left[\frac{\partial T_{[n]}}{\partial v} \frac{\partial T_{[m]}}{\partial u} - \frac{\partial T_{[n]}}{\partial u} \frac{\partial T_{[m]}}{\partial v} \right] du dv, \quad \text{if } n \neq m, \\ &= j\omega \mu, \quad \text{if } n = m; \end{aligned} \quad (52)$$

⁸ C. L. Hogan, "The Ferromagnetic Faraday Effect at Microwave Frequencies and Its Applications—The Microwave Gyrator, *Bell System Tech. J.*, Jan. 1952, p. 9.

$$Z_{[m](n)} = j\omega\mu_{uv} \iint \left[\frac{\partial T_{(n)}}{e_1} \frac{\partial T_{[m]}}{e_1} \frac{\partial T_{[m]}}{\partial u} + \frac{\partial T_{(n)}}{e_2} \frac{\partial T_{[m]}}{e_2} \frac{\partial T_{[m]}}{\partial v} \right] e_1 e_2 du dv;$$

$$Y_{[m][n]} = 0, \quad \text{if } n \neq m,$$

$$= j\omega\epsilon + \frac{\chi_{[m]}^2}{j\omega\mu_{zz}}, \quad \text{if } n = m;$$

$$Y_{[m](n)} = 0, \quad \text{all } m, n.$$

We note that $Z_{[m](n)} = -Z_{[n](m)}$; $Z_{[m][n]} = -Z_{[n][m]}$, ($n \neq m$).

In rectangular waveguides we choose cartesian coordinates; then $e_1 = e_2 = 1$, $u = x$, $v = y$ and

$$\begin{aligned} T_{(pq)} &= 1_{pq} \chi_{(pq)}^{-1} (ab)^{-1/2} \sin \frac{p\pi x}{a} \sin \frac{q\pi y}{b}, \\ T_{[st]} &= 1_{st} \chi_{[st]}^{-1} (ab)^{-1/2} \cos \frac{s\pi x}{a} \cos \frac{t\pi y}{b}, \\ \chi_{(pq)}^2 &= \chi_{[pq]}^2 = \frac{p^2 \pi^2}{a^2} + \frac{q^2 \pi^2}{b^2} \equiv \chi_{pq}^2, \end{aligned} \quad (53)$$

where $1_{pq} = 2$ if neither p nor q is equal to zero and $1_{0q} = 1_{p0} = \sqrt{2}$. Hence

$$\begin{aligned} Z_{(pq)[st]} &= j\omega\mu_{xy} \frac{1_{pq} 1_{st} \pi^2}{a^2 b^2 \chi_{pq} \chi_{st}} \\ &\times \iint \left[(b/a) s p \sin \frac{s\pi x}{a} \cos \frac{p\pi x}{a} \cos \frac{t\pi y}{b} \sin \frac{q\pi y}{b} \right. \\ &\quad \left. + (a/b) t q \cos \frac{s\pi x}{a} \sin \frac{p\pi x}{a} \sin \frac{t\pi y}{b} \cos \frac{q\pi y}{b} \right] dx dy \\ &= j\omega\mu_{xy} \frac{1_{pq} 1_{st} p q [(s/a)^2 + (t/b)^2] [1 - (-)^{s+p}] [1 - (-)^{q+t}]}{\chi_{pq} \chi_{st} (s^2 - p^2) (q^2 - t^2)}, \end{aligned} \quad (54)$$

$$\text{if } s \neq p, q \neq t,$$

$$= 0, \quad \text{if } s = p \text{ or } q = t;$$

$$Z_{[pq][st]} = j\omega\mu_{xy} \frac{1_{pq} 1_{st} (p^2 t^2 - q^2 s^2) [1 - (-)^{s+\nu}] [1 - (-)^{q+t}]}{\chi_{pq} \chi_{st} (s^2 - p^2) (q^2 - t^2) ab},$$

$$\text{if } s \neq p, q \neq t,$$

$$= 0, \quad \text{if } s = p \text{ or } q = t, \text{ but not if } s = p \text{ and } q = t,$$

$$= j\omega\mu, \quad \text{if } s = p \text{ and } q = t.$$

Some of the mutual impedances vanish; thus

$$Z_{(pq)[st]} = 0, \quad (55)$$

if either $p + s$ or $q + t$ is even. If $p + s$, as well as $q + t$, is odd,

$$Z_{(pq)[st]} = \frac{4 \cdot 1_{pq} 1_{st} j\omega\mu_{xy} pq[(s/a)^2 + (t/b)^2]}{\chi_{pq}\chi_{st}(s^2 - p^2)(q^2 - t^2)}. \quad (56)$$

Similarly

$$Z_{[pq][st]} = 0, \quad (57)$$

if either $p + s$ or $q + t$ is even, provided $p \neq s$ and $q \neq t$. If $p + s$, as well as $q + t$, is odd,

$$Z_{[pq][st]} = \frac{4 \cdot 1_{pq} 1_{st} (p^2 t^2 - q^2 s^2) j\omega\mu_{xy}}{\chi_{pq}\chi_{st}(s^2 - p^2)(q^2 - t^2)ab}. \quad (58)$$

Consider now the set of modes which includes $\text{TE}_{[10]}$. This set includes $\text{TE}_{[01]}$ modes and all the other modes which are coupled to either of these modes. Noting that there are no $\text{TM}_{(p0)}$ and $\text{TM}_{(0q)}$ modes, we obtain the following table in which those modes which do not belong to the set are marked with a bar:

$$\begin{aligned} &\text{TE}_{[10]}, \text{TE}_{[01]}, \\ &\overline{\text{TE}}_{[20]}, \overline{\text{TE}}_{[11]}, \overline{\text{TE}}_{[02]}, \overline{\text{TM}}_{(11)}, \\ &\text{TE}_{[30]}, \text{TE}_{[21]}, \text{TE}_{[12]}, \text{TE}_{[03]}, \text{TM}_{(21)}, \text{TM}_{(12)}, \\ &\overline{\text{TE}}_{[40]}, \overline{\text{TE}}_{[31]}, \overline{\text{TE}}_{[22]}, \overline{\text{TE}}_{[13]}, \overline{\text{TE}}_{[04]}, \overline{\text{TM}}_{(31)}, \overline{\text{TM}}_{(22)}, \overline{\text{TM}}_{(13)}, \end{aligned} \quad (59)$$

From the preceding equations we obtain the coupling impedances,

$$\begin{aligned} Z_{[10][01]} &= \frac{8}{\pi^2} j\omega\mu_{xy}, \quad Z_{[01][10]} = -\frac{8}{\pi^2} j\omega\mu_{xy}, \\ Z_{[10][30]} &= Z_{[30][10]} = Z_{[01][03]} = Z_{[03][01]} = 0, \\ Z_{[30][01]} &= -Z_{[01][30]} = Z_{[10][03]} = -Z_{[03][10]} = \frac{8}{3\pi^2} j\omega\mu_{xy}, \\ Z_{[21][10]} &= -Z_{[10][21]} = \frac{8\sqrt{2}}{3\pi^2} j\omega\mu_{xy}[1 + 4(b/a)^2]^{-1/2}, \\ Z_{[01][12]} &= -Z_{[12][01]} = \frac{8\sqrt{2}}{3\pi^2} j\omega\mu_{xy}[1 + 4(a/b)^2]^{-1/2}, \\ Z_{[10][12]} &= Z_{[12][10]} = Z_{[01][21]} = Z_{[21][01]} = 0, \\ Z_{[10][21]} &= -Z_{[21][10]} = \frac{16\sqrt{2}}{3\pi^2} j\omega\mu_{xy}[1 + (a/b)^2]^{-1/2}, \end{aligned} \quad (60)$$

$$Z_{[01](12)} = -Z_{(12)[01]} = \frac{16\sqrt{2}}{3\pi^2} j\omega\mu_{xy}[4 + (b/a)^2]^{-1/2},$$

$$Z_{(12)[10]} = Z_{[10](12)} = Z_{(21)[01]} = Z_{[01](21)} = 0.$$

The principal effect of the gyromagnetic medium on the $TE_{[10]}$ and $TE_{[01]}$ modes may be understood by taking into account their mutual coupling but ignoring their coupling to other modes. The equations of propagation become

$$\begin{aligned}\frac{dV_{[10]}}{dz} &= -j\omega\mu I_{[10]} - j\omega\mu_{xy}(8/\pi^2)I_{[01]}, \\ \frac{dI_{[10]}}{dz} &= -\left(j\omega\epsilon + \frac{\pi^2}{j\omega\mu_{zz}a^2}\right)V_{[10]}, \\ \frac{dV_{[01]}}{dz} &= j\omega\mu_{xy}(8/\pi^2)I_{[10]} - j\omega\mu I_{[01]}, \\ \frac{dI_{[01]}}{dz} &= -\left(j\omega\epsilon + \frac{\pi^2}{j\omega\mu_{zz}b^2}\right)V_{[01]}.\end{aligned}\tag{61}$$

For exponentially propagated waves we have

$$\begin{aligned}V_{[10]} &= \hat{V}_{[10]}e^{-j\beta z}, & V_{[01]} &= \hat{V}_{[01]}e^{-j\beta z}, \\ I_{[10]} &= \hat{I}_{[10]}e^{-j\beta z}, & I_{[01]} &= \hat{I}_{[01]}e^{-j\beta z}.\end{aligned}\tag{62}$$

When the mutual permeability is zero, we have two independent modes whose phase constants are

$$\beta_{10} = \left(\omega^2\mu\epsilon - \frac{\mu\pi^2}{\mu_{zz}a^2}\right)^{1/2}, \quad \beta_{01} = \left(\omega^2\mu\epsilon - \frac{\mu\pi^2}{\mu_{zz}b^2}\right)^{1/2}.\tag{63}$$

The phase constants of the perturbed modes may be expressed in terms of the unperturbed constants and the coefficient of coupling. When the losses are neglected, the mutual permeability is a pure imaginary. In this case it is convenient to define a *real* coupling coefficient

$$k = \frac{j8\mu_{xy}}{\pi^2\mu}.\tag{64}$$

Substituting from (62) in (61) and using (64), we find

$$\begin{aligned}\beta\hat{V}_{[10]} &= \omega\mu\hat{I}_{[10]} - j\omega\mu k\hat{I}_{[01]}, & \beta\hat{I}_{[10]} &= \left(\omega\epsilon - \frac{\pi^2}{\omega\mu_{zz}a^2}\right)\hat{V}_{[10]}, \\ \beta\hat{V}_{[01]} &= j\omega\mu k\hat{I}_{[10]} + \omega\mu\hat{I}_{[01]}, & \beta\hat{I}_{[01]} &= \left(\omega\epsilon - \frac{\pi^2}{\omega\mu_{zz}b^2}\right)\hat{V}_{[01]}.\end{aligned}\tag{65}$$

Eliminating $\hat{V}_{[10]}$ and $\hat{V}_{[01]}$, we have

$$\begin{aligned}(\beta^2 - \beta_{10}^2)\hat{I}_{[10]} &= -jk\beta_{10}^2\hat{I}_{[01]}, \\ (\beta^2 - \beta_{01}^2)\hat{I}_{[01]} &= jk\beta_{01}^2\hat{I}_{[10]}.\end{aligned}\quad (66)$$

Multiplying term by term, we obtain the characteristic equation

$$\beta^4 - (\beta_{10}^2 + \beta_{01}^2)\beta^2 + (1 - k^2)\beta_{10}^2\beta_{01}^2 = 0. \quad (67)$$

Solving, we have

$$\beta^2 = \frac{1}{2}(\beta_{10}^2 + \beta_{01}^2) \pm \frac{1}{2}[(\beta_{10}^2 - \beta_{01}^2)^2 + 4k^2\beta_{10}^2\beta_{01}^2]^{1/2}. \quad (68)$$

The effect of coupling is to increase the larger phase constant and decrease the smaller one; that is, to make the slower wave slower, and the faster wave faster.

Let us assume $a > b$; then $\beta_{10} > \beta_{01}$. Taking the upper sign in (68) and substituting in the second equation of the set (66), we have

$$\frac{\hat{I}_{[01]}}{\hat{I}_{[10]}} = \frac{jk(\beta_{01}/\beta_{10})}{p + (p^2 + k^2)^{1/2}}, \quad p = \frac{1}{2}\left(\frac{\beta_{10}}{\beta_{01}} - \frac{\beta_{01}}{\beta_{10}}\right). \quad (69)$$

From (65) and (69) we find

$$\frac{\hat{V}_{[01]}}{\hat{V}_{[10]}} = \frac{\beta_{10}^2}{\beta_{01}^2} \frac{\hat{I}_{[01]}}{\hat{I}_{[10]}} = \frac{jk(\beta_{10}/\beta_{01})}{p + (p^2 + k^2)^{1/2}}. \quad (70)$$

Hence, the ratio of the power carried in the $\text{TE}_{[01]}$ mode to that in the $\text{TE}_{[10]}$ mode is

$$\frac{P_{01}}{P_{10}} = \frac{\hat{V}_{[01]}\hat{I}_{[01]}^*}{\hat{V}_{[10]}\hat{I}_{[10]}^*} = \frac{k^2}{[p + (p^2 + k^2)^{1/2}]^2}. \quad (71)$$

This ratio increases as k increases and p decreases.

If the phase constants of the uncoupled modes are equal, then $p = 0$ and $P_{01} = P_{10}$ for all values of the coupling coefficient. In this case (68) becomes

$$\beta^2 = \beta_{10}^2(1 \pm k) \quad \text{or} \quad \beta = \beta_{10}(1 \pm k)^{1/2}. \quad (72)$$

In terms of the original constants,

$$\beta = \left[\left(\mu \pm \frac{8}{\pi^2} j\mu_{xy} \right) \left(\omega^2 \epsilon - \frac{\pi^2}{\mu_{zz} a^2} \right) \right]^{1/2}. \quad (73)$$

The cutoff frequencies of both normal modes are seen to be independent of either the transverse permeability or the mutual permeability. Since

μ_{xy} is a pure imaginary, it effectively increases the transverse permeability for one mode and decreases it for the other.

To evaluate the effect of higher order TE and TM modes on wave propagation we may substitute from (68) in all terms of the characteristic equation for telegraphist's equations except the first two diagonal terms and recalculate the β 's. Alternatively we may replace $\text{TE}_{[10]}$ and $\text{TE}_{[01]}$ modes by the normal modes just obtained, recalculate the coupling coefficients, and evaluate the effect of the mode with the greatest coupling to the modes under consideration.

Photoelectric Properties of Ionically Bombarded Silicon

By EDWIN F. KINGSBURY and RUSSELL S. OHL

(Manuscript received March 25, 1952)

In the course of investigation of the rectifying properties of silicon very interesting photoelectric properties were found. The first photo-cells were cut from bulk silicon in which a natural potential barrier was found. A typical spectral characteristic of such a cell is shown. This early work was followed by the discovery of the ionic bombardment method of producing photo active silicon surfaces. The effects of the temperature of the target and of the energy of the bombarding particles in the photoelectric properties is illustrated by characteristic curves. Relative equi-energy spectral response characteristics as a function of wavelength are illustrated. The photon efficiency as a function of wavelength of a typical cell is shown.

INTRODUCTION

Because of the importance that barriers have come to assume in the general field of semiconductors the authors have been urged to publish results of their early experiments in this field. These experiments were undertaken in the course of a search for semiconductive material suitable for use as point contact rectifiers.

Before March 1941¹ one of the writers discovered a well-defined barrier having a high degree of photovoltaic response. The barrier was found only in melts of some lots of commercially available high-purity silicon. This barrier showed a high photovoltaic sensitivity to radiation from incandescent lamps.

The existence of this natural barrier was first observed in rods cut from melts for resistivity measurements. These rods showed a high degree of photovoltaic response, were found to have a high thermoelectric coefficient, and had good rectifying properties.¹ The fact that one end of the rod developed a negative potential when illuminated or heated and that when supplied with a negative potential showed low resistance to current flow across the barrier led to the terminology of *n*-type

silicon. The material of opposite type was named *p*-type. Material of the *n*-type is now known to derive its electrical properties from the presence in the crystal lattice of electron donor impurities, for example phosphorus, while *p*-type derives its electrical properties from the presence of electron acceptor impurities, such as boron. In this paper some of the results of investigations of the natural barrier are reported; however, the photovoltaic properties of induced barriers obtained by the ionic bombardment of *p*-type silicon will be given more detail treatment.

EARLY RESULTS

The approximate location of the natural barrier found in early melts is shown in Fig. 1. The barrier was generally located in the melt perpendicular to the axis of the melting crucible or more accurately to the direction of the temperature gradient. Plates and rods containing sections of the photoactive barrier, Fig. 1a, were cut from the melt and mounted on convenient supports for laboratory tests. Fig. 1c shows a magnified section of one of these barriers.²

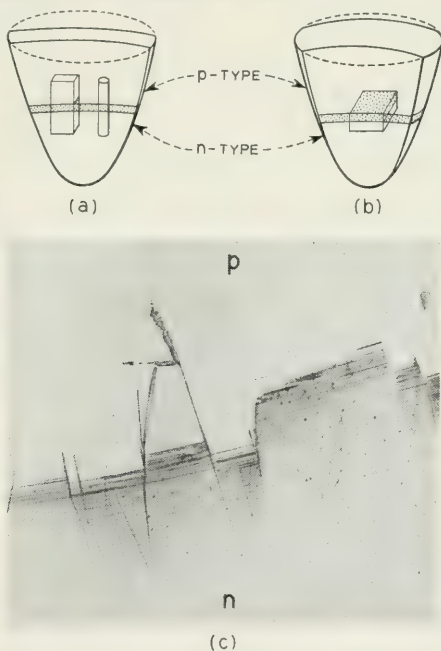
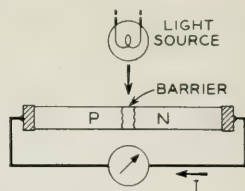
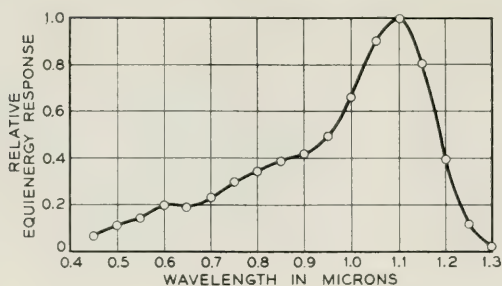
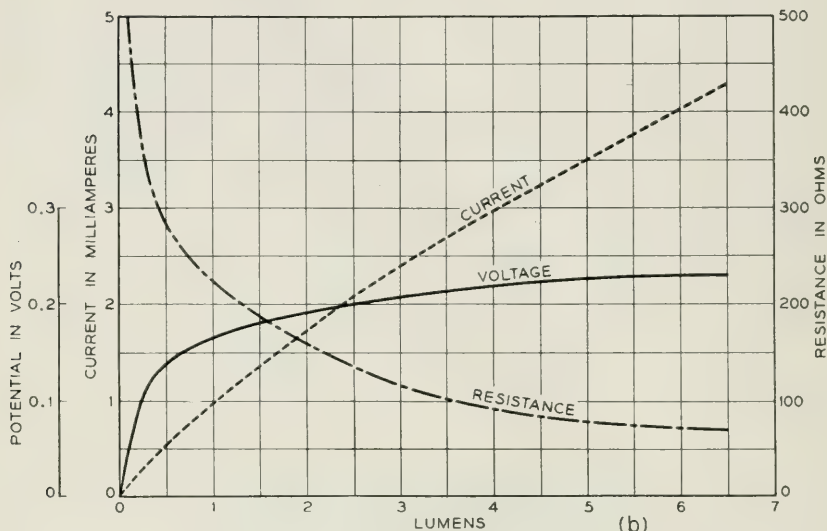


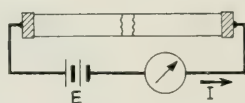
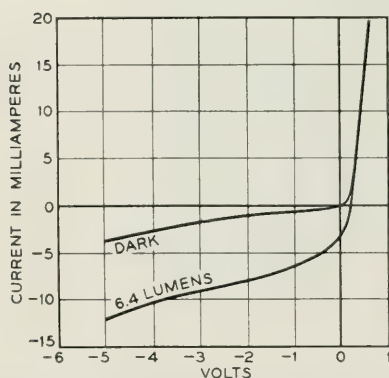
Fig. 1—(a) Drawing of melt showing position of photovoltaic barrier and photo cells with natural barrier. (b) Drawing of melt showing surface type photo cell made from natural barrier. (c) Magnified, etched section of slowly cooled silicon showing the transition between *p* and *n* silicon forming the barrier which consists of intermeshed striae of these two varieties.



(a)



(b)



(c)

Fig. 2a—Spectral response of internal barrier in silicon.
 Fig. 2b—Voltage and current photosensitivity of internal barrier in silicon.
 Fig. 2c—Rectification characteristic of internal barrier, dark and illuminated.

A typical spectral response curve of such a barrier is shown in Fig. 2a while Fig. 2b gives its open circuit voltage, short-circuit current and resistance when illuminated by a tungsten light of 2848°K color temperature. This cell resistance was taken as equal to that of an added series resistance which reduced the short-circuit photocurrent to one-half. The value so obtained is somewhat higher than the corresponding ratio of the voltage and current given in the figure. Fig. 2c gives the behavior as a rectifier in the dark and with a stated light on the barrier.

Cells whose barrier was near the surface were made by cutting close to the natural one as shown in Fig. 1b. This cut exposed large photoactive areas. Surface barrier activity was occasionally found on the top surface of some melts. These surface type cells showed a wider spectral response toward the visible than the internal barrier type. This was found to be due to the spectral absorption characteristics of the bulk silicon. A further discussion of this appears later in the paper.

These early barrier cells showed remarkable stability under severe temperature conditions. For instance, they could be heated to redness in air and quenched in water with no serious change in their characteristics. They were tested in liquid nitrogen, under water and in oil without injury. They could be illuminated with direct sunlight with no injury to their response characteristics other than the temporary effect of the increased temperature. Several of these internal barrier cells have been in use in test circuits for more than ten years with no serious change in their photoresponse properties. These observations seemed to indicate clearly that a very high degree of stability could be expected from silicon photocells.

However, there were serious practical disadvantages to the early cells. Those shown in Fig. 1a were active near the exposed barrier itself which was usually a strip along the surface about $\frac{1}{2}$ mm wide. On the other hand, the surface types as shown in Fig. 1b showed irregular responsiveness over the surface area.

From these early studies it was clear that if a good method could be found to activate large areas of silicon surfaces uniformly, cells could be made which might compete with other kinds of surface barrier type cells already available. The search for such a process resulted in the ionic bombardment method of activating silicon surfaces. Such surfaces also have desirable rectifying properties.³

METHOD OF PREPARATION

Hyper-purity silicon was used for bombardment type cells to avoid the formation of natural barriers due to minute impurities and to give

better control of the sensitivity. After being cast in a fused silica crucible, the roughly cylindrical piece was ground to a cylinder about $1\frac{1}{2}$ " diameter, a process which removed crucible contamination and gave a convenient size for slicing into wafers about 0.025" thick. The two faces were then made approximately flat and parallel after which one was left rough and the other ground and polished down to a good optical surface. In most cases the discs were then cleaned by soaking for approximately fifteen minutes in a solution of hydrofluoric acid, rinsed in distilled water and dried.

The activation consisted of exposing the polished face to a uniform beam of positive ions of helium at a pressure of 10^{-3} to 10^{-4} mm of

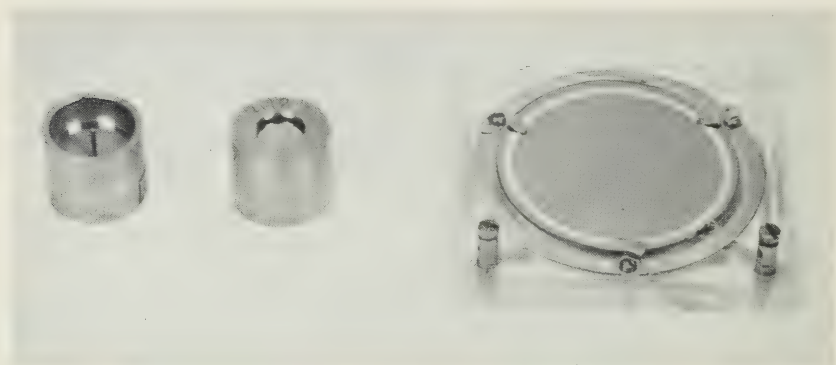


Fig. 3—Intermediate and large size photocells made by ion bombardment. Back of the intermediate also shown.

mercury. The energy of the particles used in different units ranged from 100 to 30,000 electron volts. During this bombardment the silicon surface was kept at a favorable temperature, about 395°C.

After activation, collector electrodes of evaporated rhodium were applied. Cells of three sizes have been constructed, two of which are shown in Fig. 3, the intermediate and the large one, of exposed active areas about 0.40 and 8.0 sq. cm. respectively. A small one had an area around 0.005 sq. cm. Most of the measurements reported in this paper have been made with the intermediate size.

EFFECT OF ION VELOCITY

That ion velocity has a profound effect on the voltage current characteristic of bombarded surfaces is shown in Fig. 5. These characteristics were obtained by placing a tungsten point contact under 10 gm of force,

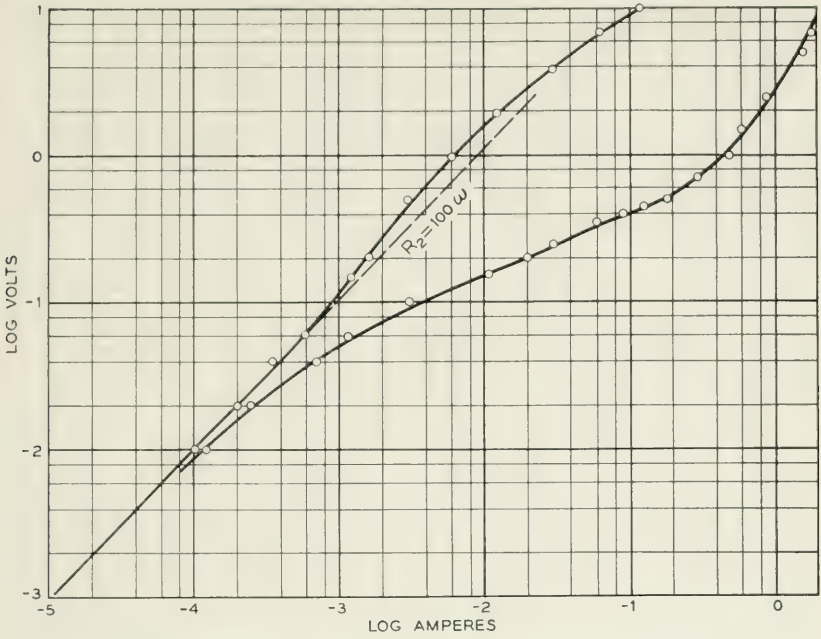


Fig. 4—Rectification characteristic of the large photocell.

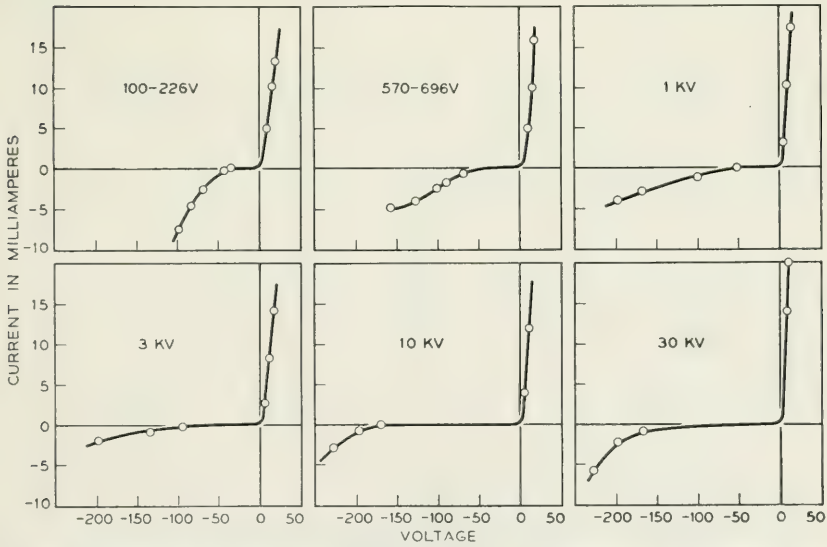


Fig. 5—Effect of bombardment voltage on the rectification of the intermediate cells.

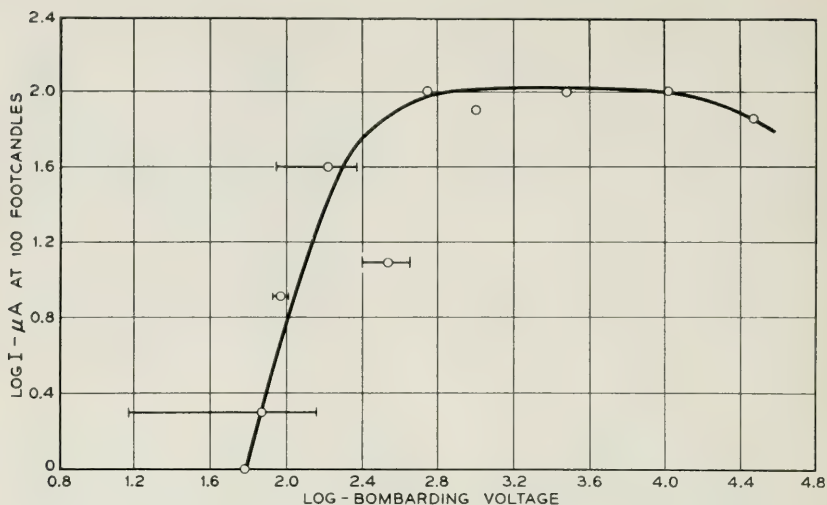


Fig. 6—Photocurrent at a constant illuminance versus the bombarding voltage.

on the photo active surfaces of the medium size cells whose spectral response is given in Fig. 8. However, in order to show the rectifying property of the barrier without the complication of a point contact, a disc of hyper-purity silicon $1\frac{1}{2}$ " in diameter and about 0.025" thick was given an optical polish on both faces. One face was bombarded with 30-kv ions.

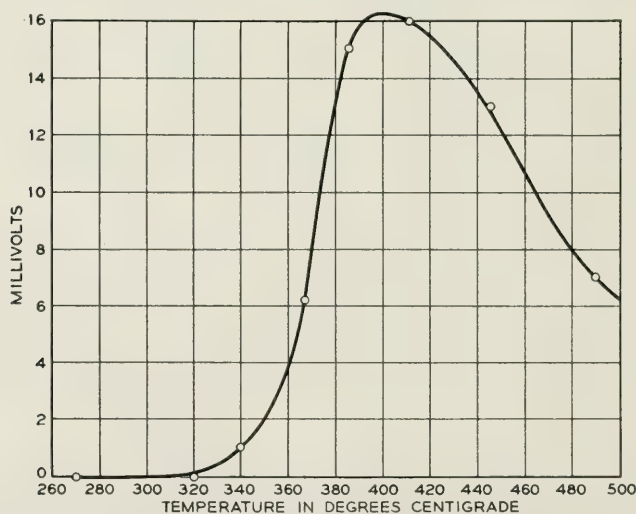


Fig. 7—Photovoltage at a constant illumination versus temperature of the bombarded silicon surface.

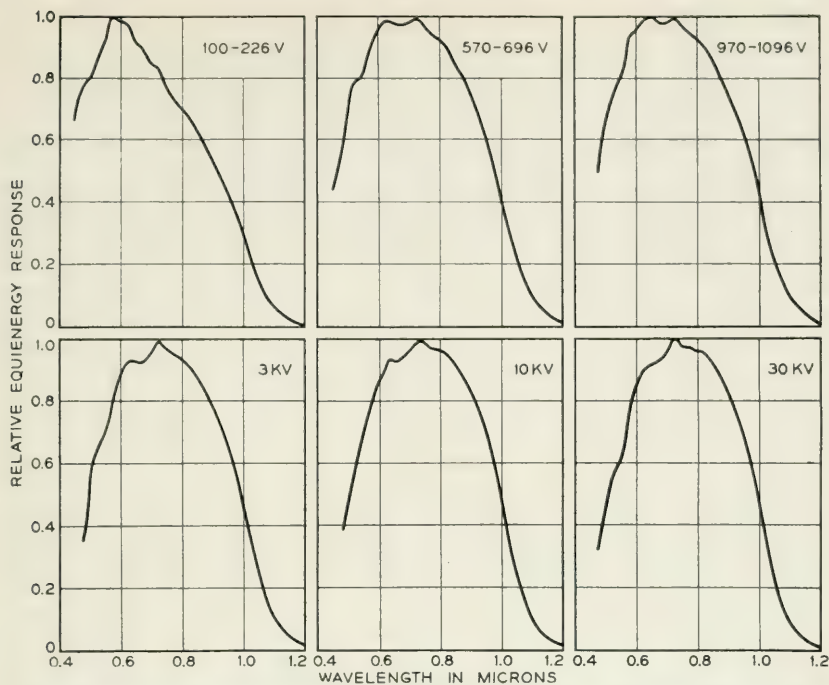


Fig. 8—Spectral response of the intermediate size cells at various bombarding voltages.

Electrodes $1\frac{1}{8}$ " in diameter of evaporated rhodium metal were applied in like manner to each surface. Contact was made to the collector electrodes by means of tin discs. Fig. 4 gives the forward and backward log voltage-log current relation of this large cell. Without bombardment such an arrangement shows ohmic conductivity so it is evident that the treatment is responsible for the development of a potential barrier beneath the surface. It is believed from the high dark resistivity of the bombarded layer that the intrinsic properties of the silicon are developed therein. Thus an intrinsic $-p$ type potential barrier is produced similar to a degree to the $n-p$ junction. One would expect the depth of this barrier to be related to the velocity of the ions. Consequently a study has been made of the effect of ion velocity on the photoelectric properties.

The photoelectric current at constant illuminance for a series of cells prepared by bombardment with ions of different energies is shown in Fig. 6. It is remarkable how quickly and completely the current sensitivity saturates at approximately 500 volts.

EFFECT OF SURFACE TEMPERATURE DURING BOMBARDMENT

In the preparation of photocells it was found that the surface temperature during bombardment had a pronounced effect on the efficiency. In order to study this effect it was necessary to determine the surface temperature of the silicon itself. Since it was impractical to measure the silicon temperature during bombardment, a calibration was made of the surface temperature in terms of the temperature of the graphite heating block. This calibration was carried out by two platinum/platinum rhodium thermocouples made of 5-mil wires. The fused thermojunction beads were held in contact with the surfaces by miniature tungsten springs. Temperature measurements with the thermojunction in contact with the silicon surface were subject to error from the slightest contamination at the point of contact. Perhaps the most difficulty was due to a reaction between the platinum and silicon at temperatures above 400°C.

The effect of surface temperature on the photoresponse is shown in Fig. 7. It is apparent that maximum sensitivity results when the target is kept at about 395°C. Perhaps by coincidence this is also the temperature at which no Hall Effect is observable in this hyper-pure material.

Cells prepared at temperatures above the critical value show lower back resistances than those prepared at the critical temperature and conversely those at temperatures below the critical value have higher back resistances but a much reduced photoresponse.

EFFECT OF TOTAL BOMBARDING CHARGE

The photoresponsiveness improves as the total bombarding charge is increased until it has reached about 600 microcoulombs per sq. cm. Further bombardment produces no appreciable improvement. In certain exploratory tests a total charge of about 9000 microcoulombs at 30 kv has been applied. Under these severe conditions the surface may show small areas having a slightly etched appearance.

No extensive tests have been made to determine the effect of the rate of application of the bombarding charge. The apparatus was designed for use at a rate of about 5 microamperes per sq. cm. It is known however, that between the limits of about 2.5 and 10 microamperes per sq. cm. the effects are subject only to the total charge or the total number of ions which strike the silicon surface.

EFFECT OF BOMBARDMENT VOLTAGE IN SPECTRAL RESPONSE

Six spectral curves are shown in Fig. 8 which illustrate the result obtained with the intermediate size cells over the bombardment voltage

range previously mentioned. The peak of the lowest voltage cell is definitely toward the blue compared with the other five whose maximum is constant at about 0.725μ .

One objective in this study was to obtain evidence relating to the depth of the barrier below the silicon surface as a function of the energy of the bombarding particles. The higher the velocity of the particles the further beneath the surface one would expect the barrier to be located and as a result there might be a shift in the spectral characteristic toward the red with increasing depth of the barrier due to the relatively greater absorption at the blue end. There is however, a selective or secondary maximum at the peak which sharpens it and nullifies the effect of the warping of the entire curve. The blue to red shift can be shown as in Fig. 9 by plotting the ratio of the responses in Fig. 8 at 0.50μ and 1.0μ . Thus at low voltage the blue to red ratio is high and decreases as the bombarding potential is raised.

In the spectral curves it will be noted that there are a number of secondary humps located near the top of the curves and extending down on the blue side. There is a strong tendency for them to occur at definite wavelengths and to be evenly spaced regardless of the bombarding voltage.

SPECTRAL MEASUREMENTS ON THE LARGE CELLS AND THE EFFECT OF MATERIAL COMPOSITION

For the large cells, two grades of silicon were used both prepared by pyrolytic reduction of SiCl_4 and called "hyper-pure". These will be

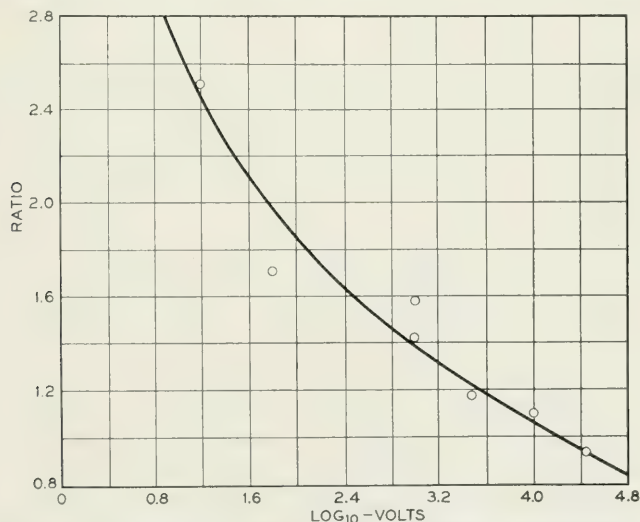


Fig. 9—Ratio of blue to near infra-red response versus bombarding voltage.

designated B and C, the former being from the same source as "silicon B" referred to in the paper by Scaff et al.² A typical analysis is given therein. The C silicon was from another source and a spectroscopic analysis indicated it was somewhat more impure than B thus agreeing with observed differences in its electrical and optical characteristics. An optical variation of considerable interest is shown in Fig. 10 where the spectral transmittance of the two grades of silicon is compared in the infra-red for polished plates each 0.0195" thick.

The transmittance of B decreases a little with increasing wavelength but C goes down much more. Both however, start to get transparent at about the same point, 1.1μ and also show corresponding absorption bands superimposed on the main curve. Briggs⁴ has compared the trans-

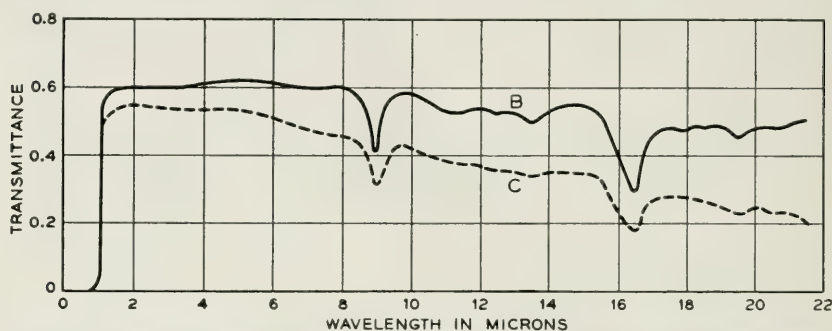


Fig. 10—Spectral transmittance of B and C grades of silicon, polished plates, 0.0195" thick.

mittance from 2μ to 12μ of the A and B silicons in Scaff's paper where the former was much more impure than the C grade. The absorption of the A silicon increased so rapidly out in the infra-red that a much thinner sample was used for the measurements than for the B material. If this difference in thickness is allowed for, the effect of impurity is very striking.

The spectral response of large area cells made of the B and C materials and bombarded with 1000-volt helium positive ions is shown in Fig. 11. The two curves are similar in shape except the one for C silicon is somewhat narrower and in addition is shifted toward the blue. Both have some of the secondary humps noted previously.

All the cells shown in this paper have indicated a long wave limit of about 1.2μ . Actually some response can usually be detected out to about 1.3μ . Measurements made some years ago on the internal barrier units also gave a limit around 1.3μ but relatively more response at 1.2μ with peaks close to 1.10μ . This difference is reasonable because light was

projected along the barrier plane and not normal to it as in the latest units, so that with the rapidly increasing transparency in this region, less infra-red radiation was lost. However, the blue was rapidly attenuated.

When illuminated by tungsten light of 2848°K color temperature, the large B cells gave 2160 microamps per lumen and the C unit 638. Correcting for a surface reflectance of 0.385, the net sensitivities would be 3510 and 1040. These measurements were made with between 4- and 5-footcandles illuminance on the cells, a region in which the response is proportional to the intensity. At much higher values of illuminance there was some falling off of response so that the effective sensitivity was a little lower. The above measurements were made on a ten ohm microammeter which is too low a resistance to affect the linearity. The intermediate cells ran approximately 3000 microamps per lumen in the most sensitive region of bombardment without correction for surface reflection and at 10- to 20-footcandles for the same tungsten lamp using a meter of 76 ohms.

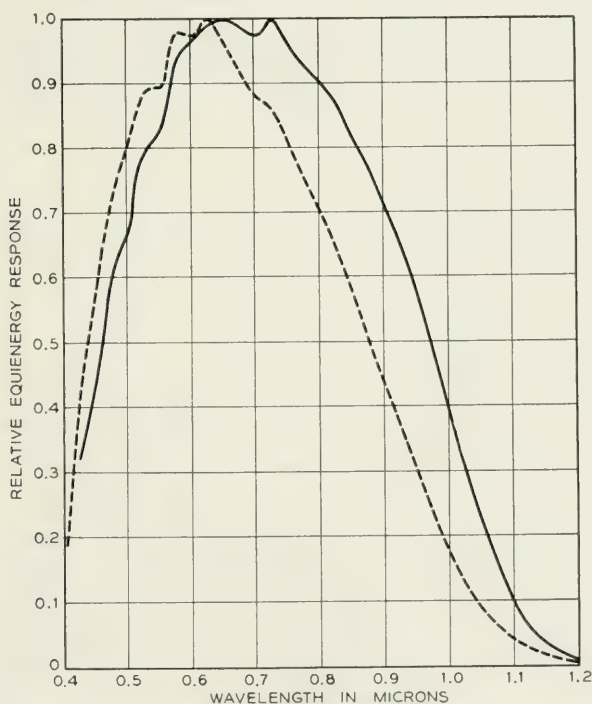


Fig. 11—Spectral response of large size photocells of B and C grades of silicon.

PHOTON EFFICIENCY

It is of interest to examine the spectral photon efficiency of a cell made by bombardment. As an example, there may be taken the 3-kv cell whose spectral response is shown in Fig. 8. When illuminated by a tungsten light of 2848°K color temperature at 10-foot candles, a sensitivity of 3090 microamps per lumen was secured. Allowing for a surface reflectance loss of 0.385, this value becomes 5020 for the radiation actually absorbed. From these data the sensitivity in microamps per microwatt at the peak 0.725 μ , calculates to be 0.388 and the photon efficiency, i.e., the electrons per photon, 0.66. Fig. 12 gives the efficiency through the spectrum. Note that the efficiency rises some on the short wave side shifting the peak of the equi-energy curve (Fig. 8) over to 0.625 μ . This increase is evident from the fact that if the equi-energy curve decreased linearly from the peak at 0.725 μ to zero at 0 μ , the photon efficiency would remain constant and equal to that at 0.725 μ . For the purpose of the above calculation, the curve in Fig. 8 has been taken as going to zero at about 0.40 μ , a fact experimentally checked. If unity is considered to be the maximum possible efficiency at any wavelength, 72 per cent of it is attained at 0.625 μ and nearly half of the spectral range is 50 per cent or higher.

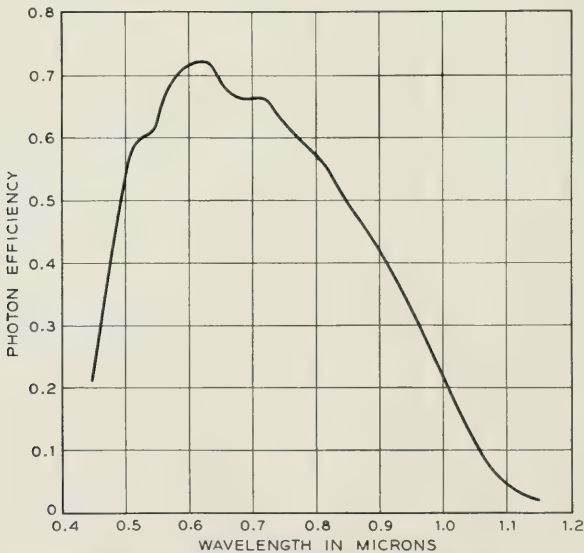


Fig. 12—Spectral photon efficiency of the 3-kv cell of Fig. 8.

CONCLUDING REMARKS

These experiments have served not only to introduce us to some of the phenomena involved in semiconductor barriers but have also yielded photo cells having desirable properties. These cells have a high degree of stability and will stand treatment ruinous to most other cells. They have a very high current sensitivity to tungsten light and daylight. They require no associated battery and can be made in large areas. Unlike the material used in many types of photo cells, silicon does not have the disadvantage of scarcity. All tests to date indicate that an indefinitely long life may be expected even under extreme illumination. Fig. 11 suggests that it may be possible to control to some extent the spectral response in the region from the deep blue into the infra-red. The long wave limit is set by the edge of the absorption characteristic.

REFERENCES

1. U. S. Patent No. 2,402,839, Filed Mar. 27, 1941.
U. S. Patent No. 2,402,662, Filed May 27, 1941.
U. S. Patent No. 2,443,542, Filed May 27, 1941.
2. J. H. Scaff, H. C. Theuerer and E. E. Schumacher; also W. G. Pfann and J. H. Scaff, *Trans. A. I. M. E.*, **185**, pp. 383-392, 1949.
3. R. S. Ohl, *Bell System Tech. J.*, Jan., 1952. Also see this paper for more details regarding the method of preparing silicon.
4. H. B. Briggs, *Phys. Rev.*, **77**, pp. 727-728, Mar. 1, 1950.

Abstracts of Bell System Technical Papers* Not Published in This Journal

Mechanical Properties of Discrete Polymer Molecules. W. O. BAKER¹, W. P. MASON¹ and J. H. HEISS¹. *J. Polymer Sci.*, **8**, pp. 129-155, Feb., 1952. (Monograph 1937).

Post-War Achievements of Bell Laboratories: II. O. E. BUCKLEY¹. *Bell Tel. Mag.*, **30**, No. 4, pp. 224-237, 1951-1952.

A Portable, Direct-Reading Microwave Noise Generator. E. L. CHINNOCK¹. *Proc. Inst. Radio Engrs.*, **40**, pp. 160-164, Feb., 1952. (Monograph 1939).

This paper discusses the factors which influenced the design of a directly calibrated portable microwave noise source, utilizing a fluorescent lamp. The variation of the noise power output and the impedance match as a function of the operating temperature are considered, and the portable unit is described.

The Quantum Theory. K. K. DARROW¹. *Sci. Am.*, **186**, pp. 47-54, Mar., 1952. (Monograph 1940).

Concerning the early years of this fundamental concept of modern physics—how Max Planck formulated it at the turn of the century and how others enlarged it up to 1923.

Performance of Ultrasonic Vitreous Silica Delay Lines. M. D. FAGAN¹. *Tele-Tech*, **11**, pp. 43-45, 138+, Mar., 1952. (Monograph 1951).

Results of tests at 10 and 60 mc with resistive terminations of 75 to 1000 ohms. Low terminating impedance values yield wide bands but involve higher insertion losses.

Phase Transition of $ND_4D_2PO_4$. B. T. MATTHIAS¹. *Phys. Rev.*, v. **85**, p. 141, Jan. 1, 1952.

Engineering Local Television Facilities and Their Operation. B. D.

* Certain of these papers are available as Bell System Monographs and may be obtained on request to the Publication Department, Bell Telephone Laboratories, Inc., 463 West Street, New York 14, N. Y. For papers available in this form, the monograph number is given in parentheses following the date of publication, and this number should be given in all requests.

¹ Bell Telephone Laboratories.

WICKLINE⁴ and J. E. FARLEY⁴. *Elec. Eng.*, **71**, pp. 252-257, Mar., 1952.

All the means of electrical communication are called into play when a city-wide coverage of an event is to be televised. How telephone and television facilities were utilized on the day that Chicago welcomed General MacArthur is explained in this article.

Echo Distortion in the FM Transmission of Frequency-Division Multiplex. W. J. ALBERSHEIM¹ and J. P. SCHAFER¹. *Proc. Inst. Radio Engrs.*, **40**, pp. 316-328, March, 1952.

The composite multiplex signals generated by frequency-division methods long standard in telephone communication can be transmitted by the new transcontinental broad-band FM radio relays. Signal intermodulation by echoes must be minimized. Such intermodulation is investigated in this paper experimentally and analytically. Two types of echoes are considered: (1) weak echoes with delays exceeding 0.1 microseconds, caused mainly by mismatched long lines; and (2) powerful echoes with delays shorter than 0.01 microseconds, caused by multipath transmission, and leading to selective fading. By use of random noise signals, the distortion is evaluated as a function of various parameters of the echo, the base-band, and the rf modulation.

Motion of a Ferromagnetic Domain Wall in Fe_3O_4 . J. K. GALT¹. *Phys. Rev.*, **85**, pp. 664-669, Feb. 15, 1952.

Experiments have been made on a sample of Fe_3O_4 cut from a single crystal in such a way that its ferromagnetic domain pattern includes an individual domain wall whose motion can be studied. This sample has a permeability which is high (about 5000) at low frequencies and drops off rapidly above 1000 cycles. A hysteresis loop and data on wall velocity vs applied field were also taken. The data are discussed in terms of recent developments in the theory of the ferromagnetic domain wall. It appears that this theory explains our data satisfactorily, and that in using it to explain our data we determine some of the fundamental magnetic constants of Fe_3O_4 . We are also able to gain some insight into domain wall motion in ferrites generally in this way.

The Drift Mobility of Electrons in Silicon. J. R. HAYNES¹ and W. C. WESTPHAL¹. *Phys. Rev.*, **85**, p. 680, Feb. 15, 1952.

Formulas for the Group Sequential Sampling of Attributes. H. L. JONES⁴. *Ann. Math. Statistics*, **23**, pp. 72-87, March, 1952.

Some Fundamental Properties of Transmission Systems. F. B. LLEWELLYN¹. *Proc. Inst. Radio Engrs.*, **40**, pp. 271-283, March, 1952.

The problem of the minimum loss in relation to the singing point is investigated for generalized transmission systems that must be stable for any combina-

¹ Bell Telephone Laboratories.

⁴ Illinois Bell Telephone Company.

tion of passive terminating impedances. It is concluded that the loss may approach zero db only in those cases where the image impedances seen at the ends of the system are purely resistive. Moreover, in such cases, the method of overcoming the transmission loss, whether by conventional repeaters or by series and shunt negative impedance loading, or otherwise, is quite immaterial to the external behavior of the system as long as the image impedances are not changed. The use of impedance-correcting networks provides one means of insuring that the phase of the image impedance of the over-all system approaches zero. General relations are derived which connect the image impedance and the image gain of an active system with its over-all performance properties.

The Arithmetic of Ménage Numbers. J. RIORDAN¹. *Duke Math. Jl.*, **19**, pp. 27-30, March, 1952.

A Recurrence Relation for Three-Line Latin Rectangles. J. RIORDAN¹. *Am. Math. Monthly*, **59**, pp. 159-162, March, 1952.

Capacitors and Communications. Inductive Coordination of Lines. A. R. WAHNER⁵ and W. E. BLOECKER². *Elec. Light and Power*, **30**, pp. 105-108, 114, March, 1952.

Although the use of capacitors on power lines has been expanding, their use has caused relatively few cases of noise on communication lines and these have been satisfactorily corrected. The causes of trouble and remedial measures were the subject of a recent, joint E.E.I.-Bell System study described here.

Book Reviews

ANTENNAS: THEORY AND PRACTICE. By Sergei A. Schelkunoff and Harald T. Friis, 639 + xxii pages, John Wiley and Sons, Inc., New York (1952). Price: \$10.00.

This is a recent addition to Wiley's Applied Mathematics Series edited by I. S. Sokolnikoff. It contains a thorough and balanced treatment of electromagnetic radiation and electrical properties of various types of antennas. In these days of rapid expansion of microwave engineering it would have been easy to neglect the older and less glamorous long-wave and short-wave antennas. The authors are to be congratulated on their impartiality. The exposition is lucid. While the entire quantitative theory of antennas is based on Maxwell's equations, unnecessary mathematics is conspicuous by its absence, and physical explanations are abundant.

The book begins with a long chapter on Physical Principles of Radiation. This chapter is almost a book within the book. It touches upon the most important ideas and problems of antenna analysis and contains a number of simple but useful formulas. Circuit and field concepts are compared, and the similarities as well as the differences between them are exhibited. Maxwell's equations are stated in a form which is particularly easy to understand. In this form, one

¹ Bell Telephone Laboratories.

² American Telephone and Telegraph Company.

⁵ Line Material Company.

equation expresses a relation between the average electric intensity tangential to a given curve and the time rate of change of the average magnetic intensity normal to a surface bounded by this curve. The other equation expresses a complementary relation. The reader will be impressed by a simple physical picture from which the authors are able to derive the expression for the radiation field of a short antenna. In this chapter they discuss the effect of heat loss and impedance mismatch on the efficiency of antennas. Among other topics will be found directive radiation and reception, large antenna arrays, horns, reflectors, and lenses.

After this extended general introduction a more detailed analysis of various problems begins. Chapter 2 is devoted to Maxwell's equations and Chapter 3 to plane waves on conductors and in free space. The main topic in Chapter 4 is the derivation of the expressions for the complete field surrounding a short antenna from Maxwell's equations. The authors have made a special effort to show the connection between this field and the oscillating charge in the antenna.

Applications of this basic theory begin with Chapter 5 devoted to directive radiation. This chapter is concerned with radiation patterns of various arrays and with calculation of radiated power. A novel method, the *method of moments* (pp. 162-195), is likely to prove valuable when spatial distributions of antenna current are complicated (as in the case of shunt-fed antennas). Chapter 6 explains methods for calculating directivities and effective areas of antennas. Some ground effects are considered briefly in Chapter 7. In Chapter 8, the discussion of current distributions in antennas made up of thin wires is particularly thorough. First, simple approximations are developed; then the effects of various factors are carefully examined. Various reciprocity and circuit equivalence theorems, so useful in antenna analysis, are collected in Chapter 9.

Beginning with Chapter 10 the general theory is applied to specific antenna types. Thus, small antennas are treated in Chapter 10; quarter-wave, half-wave and full-wave antennas in Chapter 11; general dipole antennas in Chapters 12 and 13; rhombic antennas in Chapter 14; miscellaneous types of wire antennas in Chapter 15; horn antennas in Chapter 16; slot antennas in Chapter 17; reflectors in Chapter 18; and lenses in Chapter 19.

Practical engineers will be delighted with the appendices which contain in a compact form some of the most useful information about transmission lines, dipole antennas, antenna arrays, optimum horns, and lenses. Teachers will welcome the numerous problems scattered throughout the book.

ADVANCED ANTENNA THEORY. By Sergei A. Schelkunoff, 216 + xii pages, John Wiley and Sons, Inc., New York (1952). Price: \$6.50.

This book is a recent addition to Wiley's Applied Mathematics Series edited by I. S. Sokolnikoff. It is concerned with recent advances in antenna theory and is divided into six chapters. General expressions in spherical coordinates are derived for electromagnetic fields in free space and in the presence of conducting cones and thin wires diverging from a common point. In Chapter 2 these expressions are applied to dipole antennas, vee antennas, end-fed antennas, etc. Chapter 3 gives an account of Stratton and Chu's theory of spheroidal antennas. Integral equations in antenna theory and Hallen's method of their solution are treated in Chapters 4 and 5. The book is concluded with a chapter on natural oscillations in antennas. A substantial number of problems and several interesting appendices will be found at the end.

Contributors to this Issue

SIDNEY DARLINGTON, B.S., Harvard University, 1928; B.S. in E.E., Massachusetts Institute of Technology, 1929; Ph.D., Columbia University, 1940. Bell Telephone Laboratories, 1929-. Dr. Darlington has been engaged in research in applied mathematics with emphasis on network theory.

PAUL G. EDWARDS, B.E.E., Ohio State University, 1924; E.E., Ohio State University, 1929. Western Union Telegraph Company, 1919-22; American Telephone and Telegraph Company, 1922-34; Bell Telephone Laboratories, 1934-. His main concern in the Laboratories has been with toll transmission problems, including voice frequency and carrier systems. Member of the I.R.E., A.I.E.E., Sigma Xi, Tau Beta Pi, and Eta Kappa Nu.

C. W. HARRISON, B.S. in E.E., Purdue University, 1938; M.S., Lehigh University, 1940. Bamberger Broadacasting Company, 1939-41. Bell Telephone Laboratories, 1941-. Mr. Harrison is a member of the television research group. He formerly designed radio receivers and, later, microwave relay repeaters. Member of the I.R.E.

JOHN L. HYSKO, B.S. in E.E., Cooper Union, 1921. U. S. Army, 1918-19. Bell Telephone Laboratories, 1919-. Mr. Hysko's principal activities in the Laboratories have been in the development of amplitude-modulation and frequency-shift carrier telegraph systems for land line, radio teletypewriter and submarine cable applications.

EDWIN F. KINGSBURY, B.S., Colgate University, 1910. United Gas Improvement Company, 1910-18. U. S. Army, 1918-19. Eastman Kodak Company, 1919-20. Bell Telephone Laboratories, 1920-51. Mr. Kingsbury retired in 1951 after a career which was primarily concerned with television research and development, especially that part dealing with photoelectric and electrooptical problems. Member of the Franklin Institute, the Optical Society of America, and Phi Beta Kappa; Fellow of the American Physical Society and the American Association for the Advancement of Science.

ERNEST R. KRETZMER, B.S., Worcester Polytechnic Institute, 1944; M.S., Massachusetts Institute of Technology, 1946; Sc.D., Massachusetts Institute of Technology, 1949. As a member of the Electrical Engineering Department at Massachusetts Institute of Technology, Dr. Kretzmer taught from 1944-46 and conducted research there from 1946-49. Bell Telephone Laboratories, 1949-. He works in the television research group, where he has been principally concerned with decorrelation of television signals. Member of I.R.E. and Sigma Xi.

L. R. MONTFORT, E.E., University of Virginia, 1926; American Telephone and Telegraph Company 1926-34; Bell Telephone Laboratories, 1934-. Mr. Montfort has been concerned with the engineering of carrier systems. This has included field work with new systems and field tests prior to the design of new systems. During the end of World War II and for a short time thereafter, he assisted in the engineering and testing of microwave radio systems. Member of A.I.E.E., Tau Beta Pi, Theta Tau, and Sigma Phi Epsilon.

RUSSELL S. OHL, B.S. in Electro-Chemical Engineering, Pennsylvania State College, 1918; U. S. Army, 1918 (2nd Lieutenant, Signal Corps); Vacuum tube development, Westinghouse Lamp Company, 1919-21; Instructor in Physics, University of Colorado, 1921-1922. Department of Development and Research, American Telephone and Telegraph Company, 1922-27; Bell Telephone Laboratories, 1927-. Mr. Ohl has been engaged in various exploratory phases of radio research, the results of which have led to numerous patents. For the past ten or more years he has been working on some of the problems encountered in the use of millimeter radio waves. Member of American Physical Society and Alpha Chi Sigma and Senior Member of the I.R.E.

B. M. OLIVER, B.A., Stanford University, 1935; M.S., California Institute of Technology, 1936; Ph.D., California Institute of Technology, 1939. Bell Telephone Laboratories, 1939-52. During World War II, Dr. Oliver was engaged in radar research and the rest of his employment before leaving the laboratories was in the television research group. Member of I.R.E. and Phi Beta Kappa.

WILTON T. REA, B.S., Princeton University, 1926; American Telephone and Telegraph Company, 1926-34; Bell Telephone Laboratories, 1934-. Except for the years 1941-45, when he worked on military projects. Mr. Rea has been principally concerned with telegraphy. As Tele-

graph Development Engineer, he is in charge of the development of telegraph and telephotograph systems. Senior member of I.R.E. and member of A.I.E.E. and Phi Beta Kappa.

L. C. ROBERTS, A.B., Harvard University, 1916; B.S. in E.E., Harvard University, 1919; B.S. in E.E., Massachusetts Institute of Technology, 1918; American Telephone and Telegraph Company, 1917-34; Bell Telephone Laboratories, 1934-. Mr. Roberts has been primarily concerned with the development of dc and carrier telegraph except during World War II when he worked on multichannel and single-channel radio teleprinter developments. Member of A.I.E.E.

S. A. SCHELKUNOFF, B.A., M.A. in Mathematics, The State College of Washington, 1923; Ph.D. in Mathematics, Columbia University, 1928. Engineering Department, Western Electric Company, 1923-25; Bell Telephone Laboratories, 1925-26. Department of Mathematics, State College of Washington, 1926-29. Bell Telephone Laboratories, 1929-. Dr. Schelkunoff has been engaged in mathematical research, especially in the field of electromagnetic theory.

THE BELL SYSTEM

Technical Journal

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

VOLUME XXXI

SEPTEMBER 1952

NUMBER 5

Automatic Switching for Nationwide Telephone Service

A. B. CLARK AND H. S. OSBORNE 823

Fundamental Plans for Toll Telephone Plant

J. J. PILLIOD 832

Nationwide Numbering Plan

W. H. NUNN 851

Automatic Toll Switching Systems

F. F. SHIPLEY 860

Mathematical Theory of Laminated Transmission Lines—Part I

SAMUEL P. MORGAN, JR. 883

Electrical Noise in Semiconductors

H. C. MONTGOMERY 950

Important Design Factors Influencing Reliability of Relays

J. R. FRY 976

Impedance Bridges for the Megacycle Range

H. T. WILHELM 999

Abstracts of Bell System Papers Not Published in this Journal 1013

Contributors to this Issue

1020

THE BELL SYSTEM TECHNICAL JOURNAL

ADVISORY BOARD

S. BRACKEN, *President, Western Electric Company*

F. R. KAPPEL, *Vice President, American Telephone
and Telegraph Company*

M. J. KELLY, *President, Bell Telephone Laboratories*

EDITORIAL COMMITTEE

A. J. BUSCH

F. R. LACK

W. H. DOHERTY

J. W. MCRAE

G. D. EDWARDS

W. H. NUNN

J. B. FISK

H. I. ROMNES

E. I. GREEN

H. V. SCHMIDT

R. K. HONAMAN

EDITORIAL STAFF

PHILIP C. JONES, *Editor*

M. E. STRIEBY, *Managing Editor*

R. L. SHEPHERD, *Production Editor*

THE BELL SYSTEM TECHNICAL JOURNAL is published six times a year by the American Telephone and Telegraph Company, 195 Broadway, New York 7, N. Y. Cleo F. Craig, President; Carroll O. Bickelhaupt, Secretary; Donald R. Belcher, Treasurer. Subscriptions are accepted at \$3.00 per year. Single copies are 75 cents each. The foreign postage is 65 cents per year or 11 cents per copy. Printed in U. S. A.

THE BELL SYSTEM TECHNICAL JOURNAL

VOLUME XXXI

SEPTEMBER 1952

NUMBER 5

Copyright, 1952, American Telephone and Telegraph Company

Automatic Switching for Nationwide Telephone Service

By A. B. CLARK and H. S. OSBORNE

(Manuscript received May 15, 1952)

A plan for automatic long distance switching, which will ultimately embrace the entire area of the United States and extend into Canada and perhaps Mexico, has been formulated and important steps have been taken toward its realization. The plan contemplates that when a telephone customer places a call with a long distance operator, this operator will be able to establish a connection to any desired telephone simply by playing a 10 or 11 digit code into an automatic mechanism. She will receive distinctive signals when the called telephone answers or when the telephone or the toll circuits are busy. She will completely control the establishment of the connection and will have available to her the information necessary for proper billing of the call. The plan also contemplates that telephone customers will ultimately be able to dial long distance calls themselves, wherever may be the locations of the calling and called telephones.

INTRODUCTION

Ever since the invention of the telephone 76 years ago, development work has been pressing forward both in telephone transmission and in switching. These two fields have been closely interrelated in the development of telephone service on a nationwide basis, and neither could have progressed as it has without corresponding progress in the other.

The first development of equipment for the mechanical switching of telephone lines was the local dial system to enable one customer to be

connected with another in the same town. It was a natural step to develop the equipment so that operators in nearby towns could complete toll calls through this local dial equipment. This was done first by using the local equipment and then with progressive modifications making it more and more suitable for toll.

By these means through the decades of the 20's and 30's regional networks were developed for operator toll dialing, using step-by-step types of equipment, particularly in Southern California, Connecticut and Ohio. Also many short haul toll calls in metropolitan areas were handled in connection with the panel type dial equipment which was developed for automatic switching in these areas.

Also during this period the range of customer dialing in large metropolitan areas was extended, where local service is measured by message registers, through arrangements for the multiple registration of calls for which the charge was more than one local unit.

An important feature of switching development in this period was the perfecting of "common control" switching systems for large metropolitan areas endowed with a high degree of intelligence and great reliability.¹ As will be shown, still more extensive and complicated functions must be performed by the common control systems of a nationwide automatic switching system.

Also throughout this period great advance was made in the quality and stability of long distance circuits. Telephone connections, some with as many as five circuits in tandem, were being regularly established by telephone operators with satisfactory overall transmission. The limitation was in the speed and accuracy with which multiple switches could be made by operators rather than in the overall transmission characteristics.

Several factors have worked together to bring about a big expansion of long distance telephone service. These include the great growth in the numbers of telephones in service, improvements in long distance transmission, in switching, and in methods of traffic operation. Since automatic switching becomes increasingly attractive as the traffic density increases, this large growth pointed toward the desirability of further mechanizing the switching operations.

In 1943 there was cut into service in Philadelphia the first installation of the No. 4 toll crossbar system.² This system was designed to enable general automatic switching of toll connections in and out of large metropolitan areas and had many of the capabilities necessary for nationwide switching.

The various considerations already mentioned, coupled with the suc-

cess of the No. 4 installation at Philadelphia, led to studies of the service and operating results which might be expected from a nationwide extension of automatic switching. The conclusion was reached that this would be a desirable objective of the Bell System companies and would result in a very substantial further improvement in the speed and accuracy of handling of long distance messages. Accordingly, during the next few years, a national plan was prepared and was adopted by the telephone companies.

GENERAL PLAN FOR NATIONWIDE AUTOMATIC SWITCHING

The features of this nationwide plan and the present status of its application form the subject of the three technical papers which accompany this introductory paper.^{3, 4, 5} The basic requirements to be met in the development of this plan included the following:

1. It should be suitable for the nationwide extension of automatic switching both by originating toll operators and by the customers direct.

When this work was commenced it was clear that a program leading toward general nationwide operator dialing was desirable. Subsequent developments have confirmed the wisdom of making the basic plan consistent with general nationwide customer dialing as well since it now appears that a very wide extension of this form of service will become desirable.

2. The plan must provide for satisfactory overall service between any two telephones in this country and Canada.

Under manual operation satisfactory overall service was provided for by the general toll switching plan in use since about 1930. This plan is modified to recognize the far greater speed and accuracy of automatic switching compared with manual switching. This involves also modifications of transmission design standards so that the overall connections will continue to be satisfactory.

3. The system must be designed for instantaneous service, so that delays due to lack of circuits or equipment would be very infrequent. This is necessary, both from the standpoints of service and the avoidance of tieups, particularly of the automatic switching machinery.

A trunking system must therefore be devised which will most economically meet this requirement, considering overall costs of lines, switching equipment and operation.

4. Machines must be designed for use at strategic points in the network, called "control switching points", to perform automatically the various tasks required to make the overall plan operative and economical.

5. The entire plan must be such as to provide satisfactorily for growth, for flexibility to meet changing conditions and for minimum overall costs of operation.

FUNDAMENTAL PLANS FOR TOLL PLANT

Mr. Pilliod's paper, pages 832 to 850, discusses the fundamental layout of plant for nationwide operator toll dialing. This is subject to changes from time to time with further specific studies, as is the case with all far-reaching fundamental plans of this type. The additional requirements imposed by nationwide customer dialing are still under study as will be discussed a little later.

The national toll switching plan is modified so that there may be a maximum of eight toll circuits switched together to connect any two telephones compared with the previous limit of five.⁶ In order to handle the entire traffic of the country, approximately 100 control switching points are necessary at which highly intelligent common control switching systems of the No. 4 crossbar type will be placed.

A very important feature of the layout is a trunking plan providing for a high degree of use of alternate routes. To design all of the toll circuit groups of the country for a no-delay service would be very expensive. However, taking advantage of the extreme rapidity of automatic switching and the ability to build into the machine capacity for using a large number of alternate routes, a trunking system has been devised in which only about one-sixth of the toll circuit groups of the country need be engineered on a very liberal basis. These are called final groups and are the groups to which the machine ultimately appeals if all of the more direct circuit groups are busy. These more direct circuit groups can then be engineered on a basis providing for high usage of the circuits, recognizing that when one group is busy the machine appeals to another and so on until as a last resort the final group is used.

In determining means for handling all of the toll messages with a relatively small number of control switching points, tremendous advantage was derived from modern transmission developments, particularly carrier systems which give a great economy from the concentration on a long distance route of large numbers of telephone circuits — numbers often running into the thousands. As a result, a considerable degree of circuitous routing and back hauling of circuits is economical if by these means the circuits can be concentrated on heavy routes. This in turn lends itself to a plan using a minimum of control switching points.

NATIONWIDE NUMBERING PLAN

In the previous use of automatic switching by toll operators, the operators were furnished with codes by means of which could be selected the various circuits necessary to reach the destination. These codes were dialed, followed by the local number of the called party. With this system, toll operators calling a given telephone from different remote cities would, in general, use different codes corresponding to the different circuit groups which they must select.

For nationwide toll dialing even by operators this system would have impossible complications, and for nationwide customer dialing it is clear that the code to be dialed must uniquely represent the office which serves the called telephone and that office only and not be dependent upon the route to be followed to reach it. In other words, it involves the development of what is called a destination type code. Another description of this code plan is to say that for toll dialing purposes each telephone in the country (and Canada) must have a distinctive telephone number different from that of every other telephone.

It is also clear that as a practical matter this number should be based upon the local telephone number of the customer prefixed by a minimum number of digits, following easily understood rules.

To bring this about has involved a very high order of planning. Such a plan has been perfected and forms currently the basis for the determination of the coding of all new telephone offices and for changes in office codes when these are necessary. The development of this is the subject of Mr. Nunn's paper.

CUSTOMER TOLL DIALING

When the customer is to dial long distance calls directly without assistance from any operator, two additional requirements are imposed beyond those necessary for nationwide operator dialing.

1. The customer normally is connected to a local central office but for the purpose of nationwide toll dialing he must be connected to the nationwide toll network. At present he does this by dialing a code such as '211' which connects him with the long distance operator. This procedure could be continued. However, since the customer must in any event dial 10 digits for the longest hauls to designate the called telephone, it is desirable if possible to cut out this preliminary step. That would mean modifying the local central office equipment so that it would receive the 10 digit numbers and transmit them on to the toll equipment. This is a simple undertaking for local central offices using the latest

type of local central office equipment, called No. 5 crossbar, which was designed with this in view.⁷ For older types of equipment, the job is more difficult.

2. The switching equipment must be provided with automatic means for recording all of the information necessary for charging the call. In the case of operator dialing this is now done manually by the operator.

Great advances have been made in recent years in the development of automatic message recording equipment. In 1944 there was placed in service in California the first installation in this country of automatic ticketing equipment.⁸ This equipment is associated with step-by-step local switching equipment and automatically prints for each call a ticket similar to that prepared by the operator with manual operation. In 1948 there was installed in Media, near Philadelphia, a greatly improved type of message recording equipment in which the information appears in the form of punched holes in a tape.⁹ This equipment is much more economical than the earlier system and also lends itself to the automatic preparation of toll statements or bills.

The present forms of equipment have been designed to be associated with local central offices. A careful study has been made of their field of application and of the basic plan necessary to provide for a general nationwide extension of customer dialing. This indicates that there will be a large field for automatic message accounting equipment associated with the toll network and arranged to receive orders for toll messages from a number of local dial offices. This centralized AMA equipment, as it is called, is under development and an initial installation will be made next year in Washington, D. C. In this installation the range of customer dialing will be limited and certain service features will be lacking, which it is planned to add later.

The nationwide extension of customer toll dialing involves many operating problems in addition to those relating to the design of the plant. These problems involve the extent to which customers wish to dial long distance calls, requiring 10 pulls of the dial, the accuracy of dialing, the treatment of wrong numbers, provision for giving subscribers information regarding telephone numbers in distant cities, information on charges and many other questions.

Recognizing that the best way to develop these questions is a trial, arrangements were made to open such a trial last fall at Englewood, N. J. This office is equipped with a No. 5 crossbar system so that arrangements for such a trial could readily be made there. The Englewood customers are able to dial directly any of about eleven million telephones in ten metropolitan areas scattered throughout the country, including

Boston, New York, Pittsburgh, Cleveland, Chicago and San Francisco and the Bay area.

The results of this trial have been very encouraging. Subscribers are continuing to dial over 95 per cent of all the calls which can be dialed. Errors due to wrong numbers are at a minimum and other difficulties are relatively low. In so far as this trial can answer the questions, the results are all in favor of the nationwide extension of customer dialing as the development and installation of facilities suitable for this purpose make it possible to do so.

In view of the prospect of nationwide customer dialing, fundamental plan studies are now being made by the Telephone Companies throughout the country of the whole layout of plant including the distribution of centralized automatic message accounting equipments with the future general application of this method of operation. The present indication is that the number of points at which toll operating centers will be required will be greatly reduced. This will react in important ways on the design of telephone buildings, telephone equipment installations and toll circuit routes.

AUTOMATIC TOLL SWITCHING AND ACCOUNTING EQUIPMENT

All of these plans depend upon the successful development of striking innovations in toll switching and automatic message accounting equipments. The plans in turn react upon the features to be incorporated in such equipments and upon the schedule of their development. Mr. Shipley's paper, pages 860 to 882, tells about the more important features of these equipments and the problems which are involved in their development.

CONCLUSIONS

Experience with operator toll dialing shows clearly that it provides a marked improvement in toll service. This improvement will increase as progress is made toward the full application of the nationwide automatic switching plan.

The development of long distance dialing by customers is at an early stage. The results of recent trials, however, indicate that nationwide customer dialing has service advantages and will generally be received with enthusiasm by telephone users. It is anticipated, therefore, that customer dialing will rapidly expand both on a regional and on a nationwide basis.

The service advantages of nationwide automatic switching are not

measured entirely by the increased speed and improved accuracy of connections. An important factor is the continued ability of the telephone system to meet the rapidly increasing demand for telephone service without making excessive demands on the available supply of labor. The development of local dial operation was absolutely necessary to handle the great growth of local telephoning which has taken place. Today, in many places, requirements for people for toll operations are very heavy and an increased amount of automatic toll switching is becoming more and more necessary to make possible handling the rapidly increasing number of long distance telephone messages.

With this development there has been a marked increase of employment. The Bell Companies today employ 244,000 operators compared with 131,000 in 1941. They have also employed many people to build and install about 300-million dollars worth of toll dialing equipment, to construct places to house it, maintain it and carry out operating rearrangements.

With respect to the future, even with the nationwide automatic switching plan in full operation and the local central offices arranged to permit customer dialing, there will still be a large amount of work for operators. They will be required to handle information and assistance traffic, person-to-person calls, collect calls and other classes of calls which do not lend themselves to customer handling, as well as any individual calls which the customers may not wish to dial themselves.

The Bell Companies have necessarily taken the lead in planning and applying these new developments. The plans, however, are all laid in such a way as to include telephone users in Independent Telephone Company offices. The Independent Companies are being kept fully informed of these plans as they develop and are participating, as the development of their own plant makes it practicable and desirable, in extending the benefits of the new forms of operation to their own customers.

This long-term development has required the very close cooperation of all parts of the Bell System - American Telephone and Telegraph Company General Department, Bell Telephone Laboratories, Western Electric Company, Long Lines and all of the Bell Operating Companies. Each installation of equipment and circuits and each operation is a part of a nationwide system and must be closely coordinated. The close interrelation and working together of the various parts of the Bell Telephone System, research and development, manufacturing, engineering and operating are necessary for the effective planning and execution of this tremendous project.

BIBLIOGRAPHY

1. F. J. Scudder and J. N. Reynolds, "Crossbar Dial Telephone Switching System," *A.I.E.E. Transactions*, **58**, pp. 179-192, 1939.
2. L. G. Abraham, A. J. Busch and F. F. Shipley, "Crossbar Toll Switching System," *A.I.E.E. Transactions*, **63**, pp. 302-309, 1944.
3. J. J. Pilliod, "Fundamental Plans for Toll Telephone Plant." Page 832 of this issue.
4. W. H. Nunn, "Nationwide Numbering Plan." Page 851 of this issue.
5. F. F. Shipley, "Automatic Toll Switching Systems." Page 860 of this issue.
6. H. S. Osborne, "A General Switching Plan for Telephone Toll Service," *A.I.E.E. Transactions*, **49**, pp. 1549-1557, 1930.
7. F. A. Korn and J. G. Ferguson, "No. 5 Crossbar Dial Telephone Switching System," *A.I.E.E. Transactions*, **69**, Part 1, pp. 244-254, 1950.
8. O. A. Friend, "Automatic Ticketing of Telephone Calls," *A.I.E.E. Transactions*, **63**, pp. 81-88, 1944.
9. John Meszar, "Fundamentals of the Automatic Telephone Message Accounting System," *A.I.E.E. Transactions*, **69**, Part 1, pp. 255-269, 1950.

Fundamental Plans for Toll Telephone Plant

By J. J. PILLIOD

(Manuscript received May 15, 1952)

This paper covers the general switching plan and fundamental plant layout proposed for handling telephone toll messages throughout the United States and Canada using automatic toll switching.

There has been rapid growth in the number of telephones and in the volume of toll traffic, particularly long haul. Toll facilities are provided under fundamental plans, an essential part of which is a toll switching plan for setting up connections quickly between any two telephones. The introduction of mechanical operation and the general improvement in the transmission performance of the communication plant over a period of years make the introduction of certain modifications in the fundamental plans possible and advantageous at this time. The important new features and the service improvements which are provided by the proposed plans are outlined in this paper. The principal types and characteristics of circuit facilities available for use in the intertoll network are also described.

GENERAL ASPECTS OF TOLL SWITCHING PROBLEMS

Switching plans providing for the systematic routing of toll telephone traffic have been employed by the communication industry for many years. These plans have contributed directly to the high quality of long distance telephone service enjoyed by the public in the United States and Canada. This generally excellent service is the result of the cooperative work of many organizations including the Bell Operating Companies, many independent connecting Companies and others in the United States as well as in adjoining countries. The techniques employed today reflect a great amount of research and engineering and improvements in manufacturing skill and in construction, maintenance and operating methods developed over a period of many years.

Throughout the United States and Canada there are approximately 20,000 different places – cities, towns, and villages – that serve as toll

connecting points. The telephone offices in each of these places have access through the toll network to practically all of the 50,000,000 telephones in the United States and Canada and also to most of the telephones in the rest of the world. Currently the Bell Operating Companies are handling toll calls at an average rate of over 7,000,000 during a business day. The many millions of different connection possibilities which this number of calls involves require a definite and comprehensive switching plan.

Whenever practicable and economical direct circuits are used to handle toll message traffic between two given points. Much of the traffic in the country is handled this way. However, a substantial volume of business, about 20 per cent, is handled as a matter of economy, by switching toll circuits together. Although the volume of traffic between different points may vary over a wide range, it is nevertheless important that adequate service be provided for all possible connections. For example, there are about 110 circuits from Chicago terminating in the toll office serving Minneapolis and St. Paul. These handle about 5500 calls per day. On the other hand, only a few calls a year may be involved between some point in Western Minnesota and a point in Florida. The switching plan described in this paper is devised for the purpose of efficiently and effectively establishing connections between any two points regardless of their separation and regardless of whether traffic volume be a few calls per year or many calls per hour.

ELEMENTS OF THE PROBLEM

In order to illustrate the problem a specific example may be useful. Fig. 1 is a map of Wisconsin and Minnesota on which nearly 1200 circles indicate points at which exchange facilities may be connected to the toll network. The extent of the coverage in this area is typical of that found throughout the country.

The 150 odd larger circles represent existing offices known as "toll centers" – that is, places where operators record toll calls and perform other operations necessary to establish toll connections. These places have switching arrangements of various types depending on how they fit into the switching plan. Some may operate as control switching points in the nationwide plan as described later.

More than 1,000 smaller circles on the map represent "tributaries" – that is, towns where little or no toll operating is done. Toll connections to and from these points are completed at the toll centers which in general do the toll operating required.

In the United States and Canada as a whole, there are approximately 2,600 toll centers. The remainder of the toll connecting points—about 17,500—are tributaries.

Fig. 2 gives an idea of the variety and complexity of the network of circuit groups required to interconnect the toll centers in one area. Here each line represents a group of circuits, known as "intertoll trunks," between two toll centers. Each group may contain anywhere from one to several dozen trunks. The location of the lines on the map is unrelated to the geographical routing of the trunks, and only a part of the circuit groups are shown. To get a complete picture one should visualize that a cluster of relatively short circuit groups radiates from each toll center to its tributaries, of which there may be up to 15 or more.

Physically, the plant consists of a network of open wire lines, cables and radio systems. On these, voice frequency or carrier operation is employed in each section as required to provide the necessary intertoll trunks. The routes of the lines in Minnesota and Wisconsin are shown by Fig. 3. In this area there are no radio routes carrying telephone circuits, but a radio system between Chicago and Minneapolis is in the planning stage.

Areas like Wisconsin and Minnesota must, of course, be connected together, and Fig. 4 shows the major Bell System toll routes that accomplish this. On a map of this kind it is not possible to include anything like the detail shown in Fig. 3. One must visualize, therefore, that each state contains a network of routes generally comparable to those shown for Wisconsin and Minnesota.

This then represents the interconnection problem to be met by an orderly switching plan that will provide efficient, reliable and fast toll telephone service between any two points.

EARLIER TOLL SWITCHING PLANS

Very early in the telephone industry it became evident that: (1) There must be a plan for connecting circuits together. (2) Switching centers with suitable equipment must be established in accordance with this plan. (3) Trunks must be provided in adequate numbers to connect every place to one or more switching centers and to interconnect the switching centers. (4) All this must be done in a way that makes it possible to provide good service at reasonable cost.

As time went on, early plans crystallized into what became known as the General Toll Switching Plan. A paper presented at the summer convention of the A.I.E.E. in Toronto in 1930 by Dr. H. S. Osborne outlined

the principles of this comprehensive plan for handling telephone toll traffic in the United States and Eastern Canada.¹ It involved two classes of major switching centers – Regional Centers and Primary Outlets – and some classes of less important centers. It also set up methods of designing toll trunks to give adequate transmission efficiency on all possible toll connections. In use for the last two decades this basic plan has been of great value in accommodating the tremendous growth of telephone toll business during this period.

SWITCHING PLAN FOR NATIONWIDE TOLL DIALING

The earlier general switching plan was based on manual switching and on a toll plant made up for the most part, of voice frequency circuits. The probability of operating irregularities and delays increases with the number of manual switches in tandem. Likewise, the transmission problem of operating many voice frequency trunks in tandem was so formidable that the number of intertoll trunks in tandem had to be limited to five. In practice, switching was avoided where practicable and economical.

Impact of Mechanization and Improved Transmission Facilities

On the other hand, mechanical switching is very fast and is designed to be practically free of operating irregularities. Delays can be minimized by fast switching to alternate routes. Also, in the last two decades the use of carrier has grown from a relatively minor place in the toll plant to the point where it is now commonplace.² Carrier provides superior transmission performance. Limitations on switching are thus greatly reduced and economies are achieved under many conditions.

In addition, mechanization of local switching systems has proceeded rapidly. With mechanized toll switching, it is becoming possible to establish many toll connections with only a single toll operator and in some cases by customer dialing, without the assistance of any operator.^{3, 4}

Along with these developments has come tremendous growth in traffic. Since 1930 toll messages in the Bell Operating Companies and the Bell Telephone Company of Canada have more than trebled, growing from an annual volume of about 650 million to about 2 billion. Intertoll trunks over 25 miles in length have increased in number from about 28,000 to about 100,000. This continuing growth in traffic volume has required a large scale development of plant facilities and has permitted a more

extensive use of carrier than would have been practicable with a slower rate of growth.

Consideration of these factors which offer an opportunity to improve service has led to the gradual reorientation of the fundamental plans for the intertoll trunk plant which is now under way.

The New General Toll Switching Plan

Mechanization of switching and the use of improved transmission instrumentalities permits the design of the switching plant to be controlled primarily by the balance between the costs of transmission facilities and of switching facilities.

The new general toll switching plan contemplates as many as eight intertoll trunks in tandem on the most complex connections to be established. These eight trunks can be interconnected at switching points as described later. The plan further contemplates that wherever possible, the traffic will by-pass intermediate switching points. The number of switches that can be avoided depends on the volume of traffic between the two points concerned and on the traffic load at the time the connection is established.

The proposed plan provides a systematic grouping of switching points. Under this arrangement, each ordinary Toll Center (TC) serves a cluster of nearby tributary points and has trunks to a "home" Primary Outlet (PO) which serves a cluster of toll centers. In some cases it appears practicable to utilize a simplified switching system at a PO, and in order to distinguish this type of center it has been designated a Tandem Outlet (TO). In turn, each PO or TO has trunks to a "home" Sectional Center (SC) which serves a section of the country varying in size from part of a state to all of several states depending on the density of the population. Similarly, The United States and Canada are divided into nine regions, each having a Regional Center (RC) serving as a central switching point for all sectional centers in the region. One of these RC's (St. Louis) is termed the National Center (NC). All of the higher orders of switching centers also act in the capacity of each of the lower centers. For example, any specific SC also acts as a PO and as a TC.

This arrangement is illustrated in Fig. 5, which covers approximately the same area as Fig. 1, portraying the toll connecting routes. Hibbing, Minnesota, is shown as a representative toll center with the tributaries it serves. It is in the service area of the Duluth Tandem Outlet, the approximate boundaries of which are indicated. Duluth lies in the Minneapolis "section," which includes a large portion of Minnesota, and is

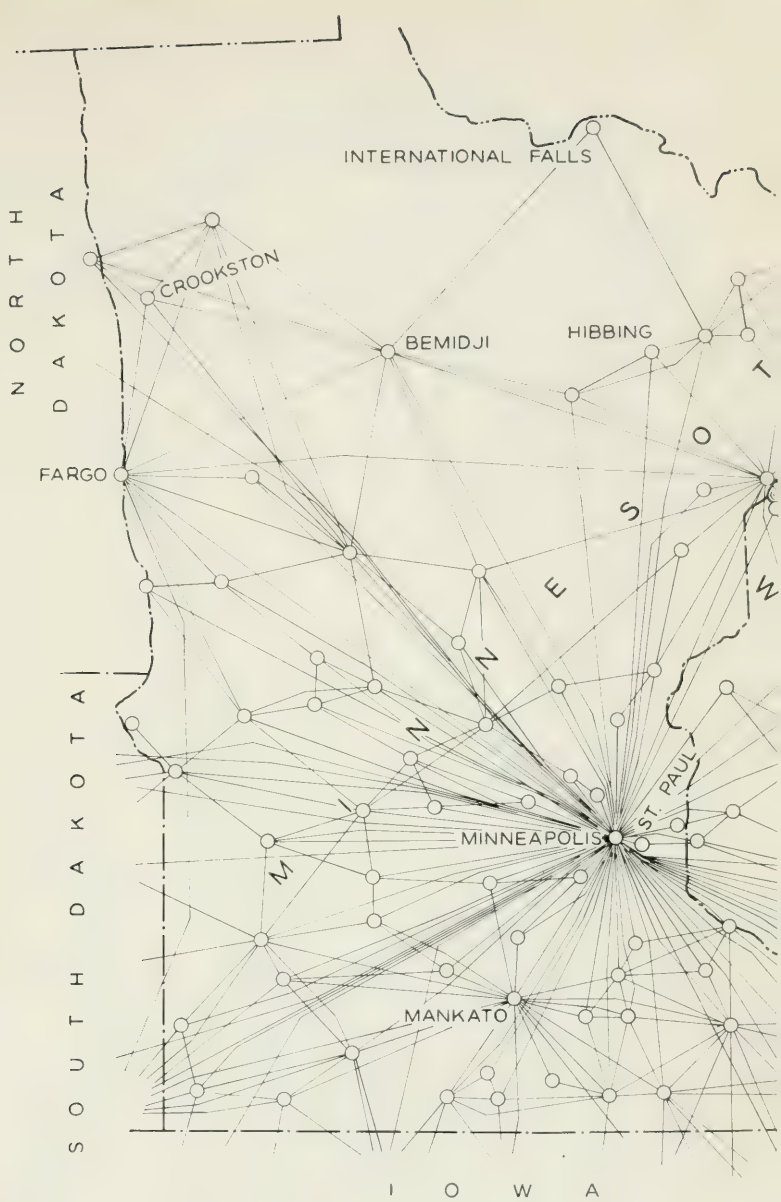


Fig. 2—Principal intertoll tru

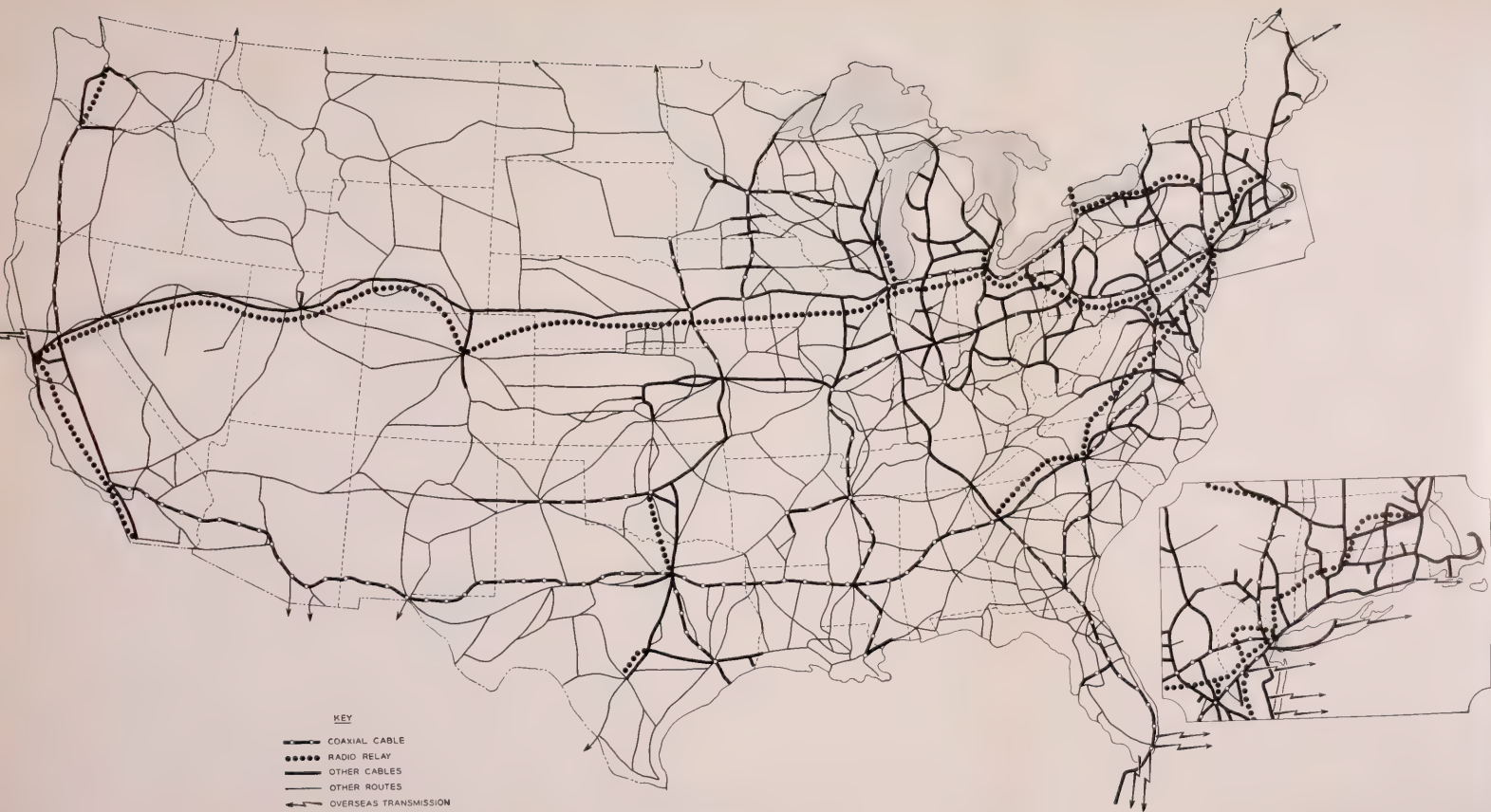


Fig. 4—Principal toll routes of the Bell System.

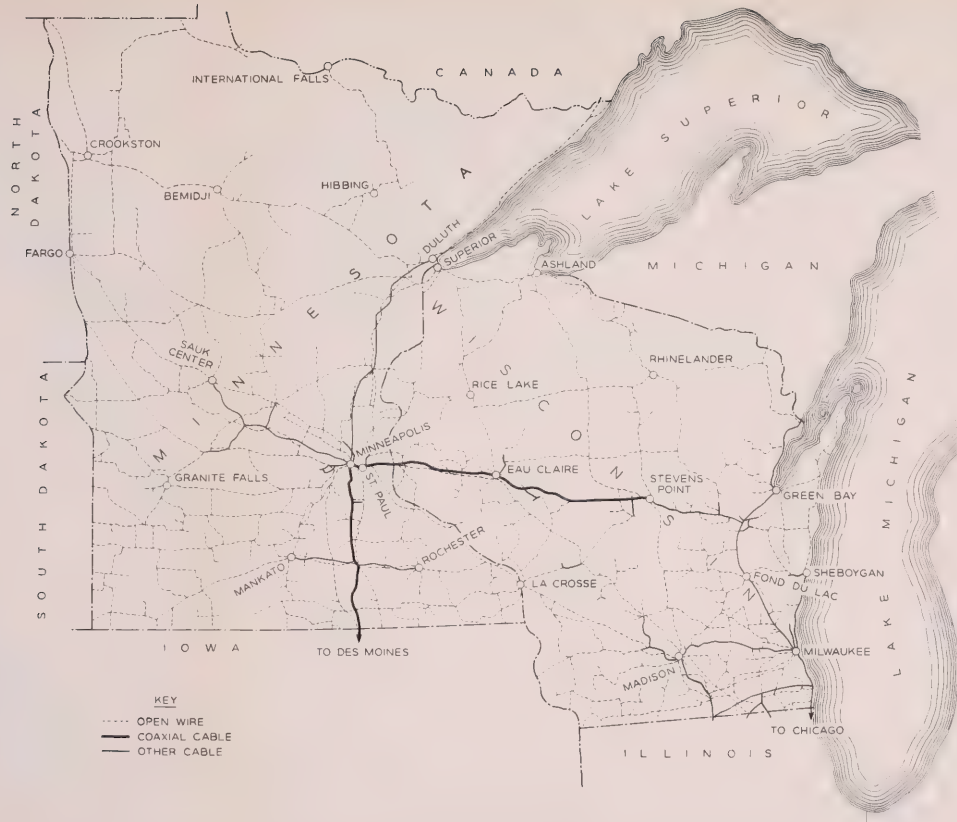


Fig. 3—Bell System telephone toll routes in Minnesota and Wisconsin.



nk groups in Minnesota and Wisconsin.

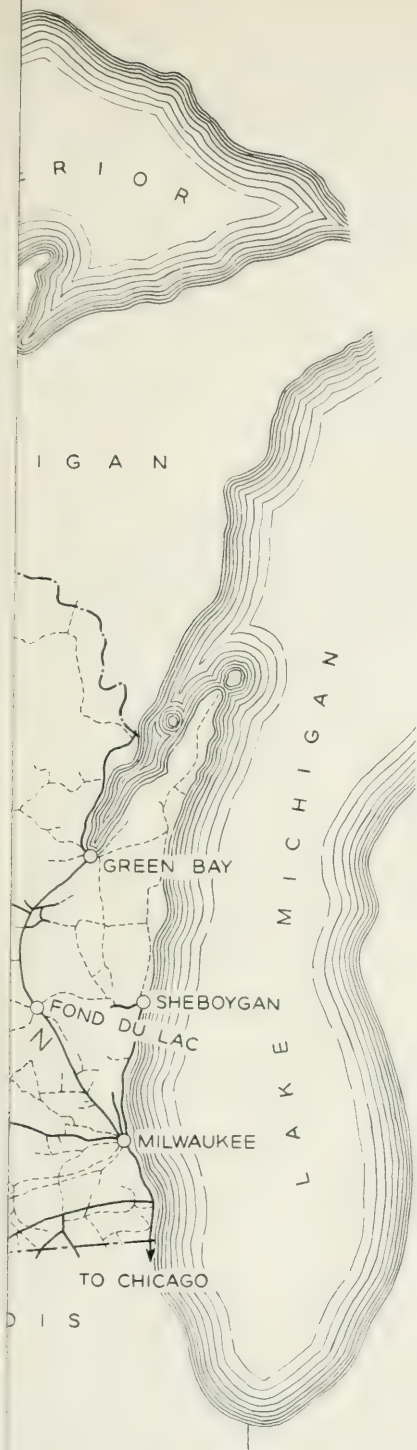




Fig. 1—Toll centers and tributaries in Minnesota and Wisconsin.

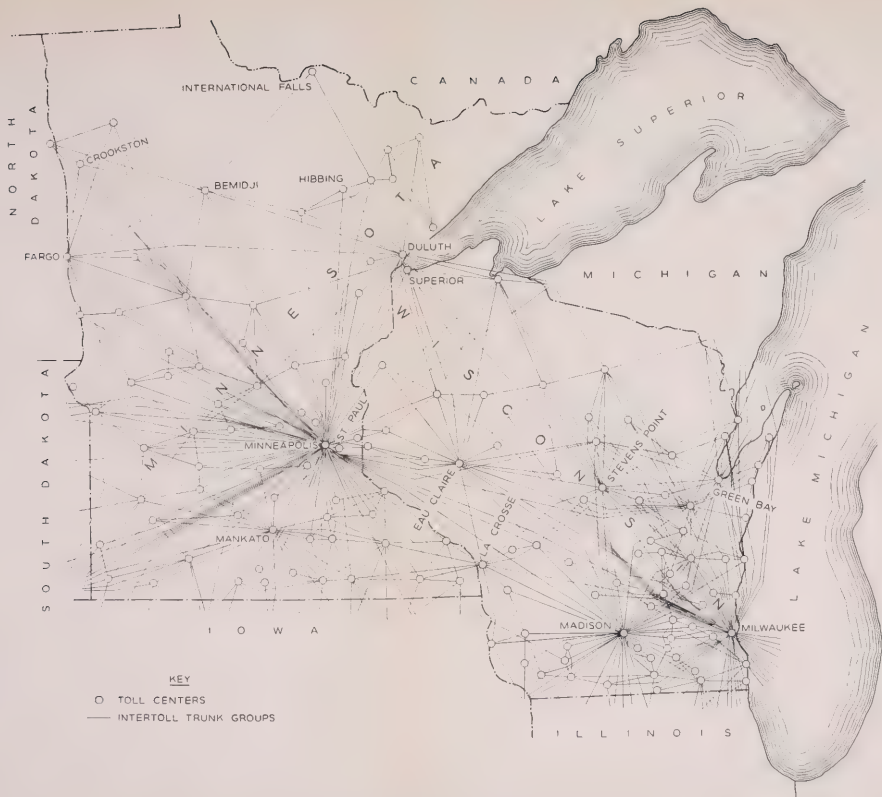


Fig. 2—Principal intertoll trunk groups in Minnesota and Wisconsin.



in turn one segment of the Chicago "region" which serves a somewhat larger area than shown by Fig. 5.

Under this arrangement, toll calls between two tributaries in the Hibbing toll center area can be completed by switching at the toll center. In a similar manner, any two points within the Duluth tandem outlet area can be served by switching at Duluth. The same treatment also applies for connections between any two points in the same sectional center area or in the same regional center area. For example, a connection from Hibbing to any point within the Chicago region (which involves more than six states as shown in Fig. 7) requires no more intertoll links than Hibbing to Duluth, Duluth to Minneapolis and Minneapolis to Chicago, and a corresponding number of links on through to another sectional center, and primary or tandem outlet to the toll center destination. Circuits between the toll center and tributaries are not referred

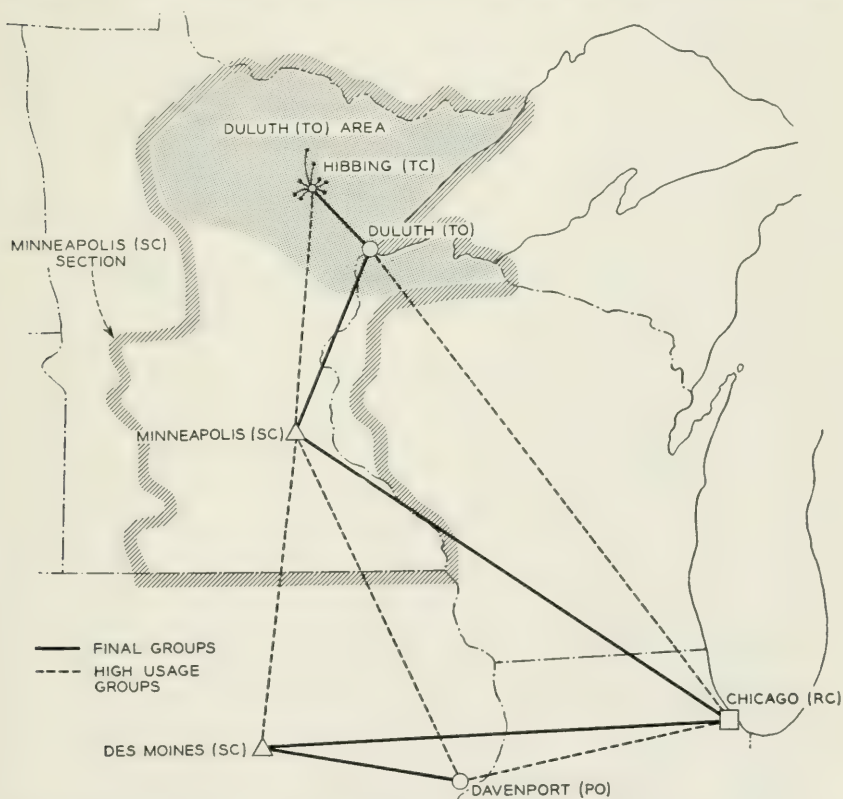


Fig. 5—Intertoll trunks between Davenport, Iowa and Hibbing, Minnesota, showing alternate routing possibilities.

to as intertoll trunks or lines but are classed as toll connecting trunks.

Where the volume of traffic warrants, direct circuits may be provided to by-pass the intermediate switching points included in the preceding example. Once such direct circuit groups have been established, it is economical and advantageous from a switching standpoint to take advantage of their existence, using routes that involve a minimum number of switches. The basic routing plan is used when the more direct circuit combinations are busy.

These routing arrangements contemplate the application of "high usage" and "final" trunk groups as an integral part of the plan. The "high usage" groups are direct groups which by-pass the higher order switching points wherever the routing of the call permits. These "high usage" groups can be engineered to carry high loads per circuit, with an adequate number of circuits in the "final" groups to take care of prac-

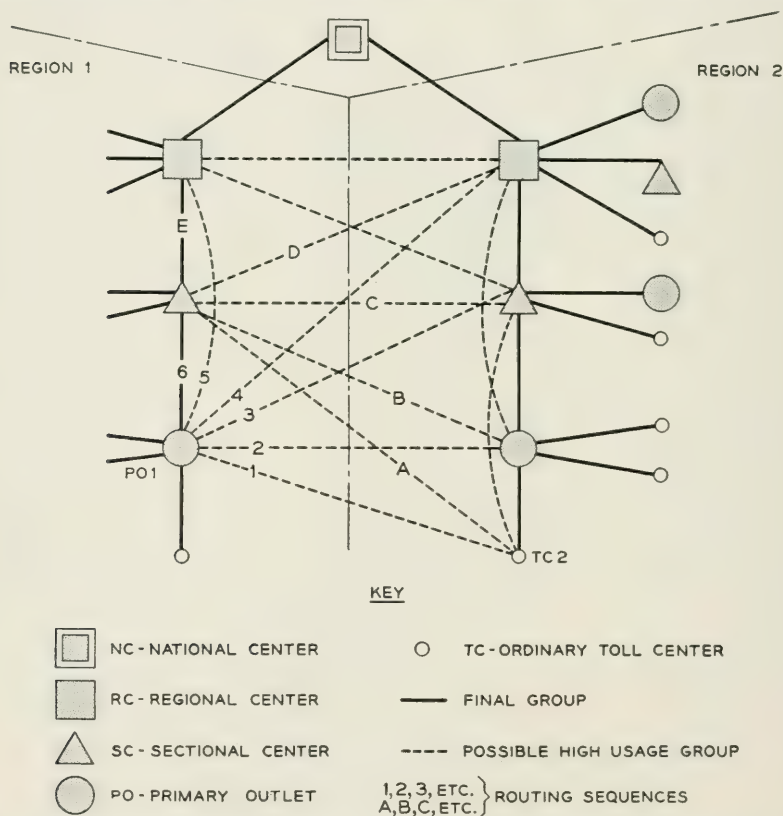


Fig. 6—Illustration of intertoll routing pattern between two regions.



Fig. 7—Tentative locations of control switching points in United States and Canada.

tically all overflows from the high usage groups during the heavy traffic periods. The "high usage" and "final" groups which could be used for routing calls between Hibbing, Minnesota and Davenport, Iowa are shown by Fig. 5.

Generalization of the Toll Switching Plan

The generalization of the arrangements discussed for the Chicago region is illustrated in Fig. 6. This shows diagrammatically all types of switching points in two regions and also indicates the relative position occupied by the National Center in the switching plan. On this chart, the solid lines represent the "final groups" of trunks, and the dotted lines represent "high usage" trunks. Examination of this chart will indicate that the mechanical switching system need perform only relatively simple toll switching operations at the toll centers. At other points the system must attempt to complete the call over the most favorable routes, in planned sequence, until the "final" route is selected.

For example, from a given primary outlet such as PO1 on a call destined for a toll center in the other region such as TC2, the switching equipment would attempt to complete the call, in sequence over the routes marked 1 to 6.

Should Route 6, which is the "final" route, be selected because all of the trunks in the "high usage" groups marked 1 to 5 were busy at the time, the switching equipment at the SC would in turn try routes marked A, B, C, etc., in attempting to complete the call. A fairly complete pattern of circuit groups is indicated in this illustration. Depending on the relative locations of the points concerned and the traffic load requirements, certain of the "high usage" groups shown may not exist. It is expected, however, that most TC's will have high usage groups to points other than their "home" PO's. Also each PO can be expected to have high usage groups to sectional centers other than its "home" SC. All regional centers will be interconnected with direct trunks, regardless of geographical location.

Control Switching Points

Because of rapid and complex switching operations required by the automatic equipment at PO's and higher order switching points, (SC's, RC's and the NC) these switching centers are called Control Switching Points (CSP's).

As covered by a companion paper,⁵ the switching equipment required at the CSP's is quite complex. This equipment must have a high degree



NOTE: THE SHORT LINE FROM EACH PO AND
SC POINTS TOWARD ITS HOME CSP

RIES

tically all overflows from the high usage groups during the heavy traffic periods. The "high usage" and "final" groups which could be used for routing calls between Hibbing, Minnesota and Davenport, Iowa are shown by Fig. 5.

Generalization of the Toll Switching Plan

The generalization of the arrangements discussed for the Chicago region is illustrated in Fig. 6. This shows diagrammatically all types of switching points in two regions and also indicates the relative position occupied by the National Center in the switching plan. On this chart, the solid lines represent the "final groups" of trunks, and the dotted lines represent "high usage" trunks. Examination of this chart will indicate that the mechanical switching system need perform only relatively simple toll switching operations at the toll centers. At other points the system must attempt to complete the call over the most favorable routes, in planned sequence, until the "final" route is selected.

For example, from a given primary outlet such as PO1 on a call destined for a toll center in the other region such as TC2, the switching equipment would attempt to complete the call, in sequence over the routes marked 1 to 6.

Should Route 6, which is the "final" route, be selected because all of the trunks in the "high usage" groups marked 1 to 5 were busy at the time, the switching equipment at the SC would in turn try routes marked A, B, C, etc., in attempting to complete the call. A fairly complete pattern of circuit groups is indicated in this illustration. Depending on the relative locations of the points concerned and the traffic load requirements, certain of the "high usage" groups shown may not exist. It is expected, however, that most TC's will have high usage groups to points other than their "home" PO's. Also each PO can be expected to have high usage groups to sectional centers other than its "home" SC. All regional centers will be interconnected with direct trunks, regardless of geographical location.

Control Switching Points

Because of rapid and complex switching operations required by the automatic equipment at PO's and higher order switching points, (SC's, RC's and the NC) these switching centers are called Control Switching Points (CSP's).

As covered by a companion paper,⁵ the switching equipment required at the CSP's is quite complex. This equipment must have a high degree

of built-in capability to perform quickly the circuit selection work associated with the alternate routing features of the switching plan. In addition, to help provide the transmission margins needed for satisfactory operation of the plan as contemplated, it must be arranged to connect circuits on a four-wire basis rather than on a two-wire basis, the latter being the arrangement used at most toll centers. The switching equipment at a CSP must not only provide for connecting one toll circuit to another; it must also perform the very important function of tying the toll networks which serve limited local areas together so that collectively they work as a smoothly functioning nationwide system. This becomes practicable when there is coordination between the design of the individual limited networks and the design of the overall system.

The location of control switching points indicated by the nationwide plan is shown in Fig. 7. This also indicates the home switching center of higher order associated with each switching point. As the number of CSP's increases, the cost of the toll circuit plant decreases because each CSP can then be located closer to the cluster of ordinary toll centers which it serves. However, because of the cost of the CSP equipment, it is necessary to weigh the cost of circuit facilities with the equipment costs in a way that will result in the minimum overall cost. Certain of the smaller Primary Outlets are being studied with the view of reclassifying them as Tandem Outlets (TO's). A Tandem Outlet occupies the same relative position in the switching plan as a Primary Outlet but is not a control switching point. The switching equipment employed is less complex than that used at control switching points and therefore provides for only limited alternate routing and does not have all the advantages of four-wire transmission.

Effects of Customer and Operator Toll Dialing

Customer dialing of short-haul toll calls has been in use, particularly in metropolitan areas, for some years. A trial of long-haul customer dialing over the intertoll trunk network and through the switching equipment provided for operator toll dialing was instituted at Englewood, New Jersey, in the Fall of 1951. The local equipment includes automatic message accounting and permits Englewood customers to dial directly to about eleven million telephones in ten metropolitan areas across the country. A trial installation of customer toll dialing, utilizing automatic message accounting equipment on a centralized basis rather than at each local office, is planned for Washington, D. C., in the Fall of 1953. Initially customers will dial toll calls within the Washington metropolitan

area and to such points as Baltimore and Annapolis. The favorable results and general acceptance of the trial at Englewood indicate extensive application of customer dialing of toll calls as conditions warrant.

The general introduction of customer toll dialing as this becomes desirable will affect the number and location of ordinary toll centers since calls handled by operators may be limited to assistance calls and to person-to-person, collect and others which cannot be customer dialed. Indications are that toll operation for a number of smaller centers can be combined as the local service is converted to dial operation with operator toll dialing.

Studies now in progress indicate that the number of toll centers may be reduced by one half or more over a period of years in many areas.

Reactions on Toll Plant Layout

The expanded general toll switching plan for nationwide dialing contemplates a degree of alternate routing far in excess of that used with the former switching plan designed for manual operation. This change along with the reduction in toll centers will have a marked effect on the normal flow of many traffic items through the intertoll network. As a result the arrangement of the present intertoll trunks will be significantly modified both in number, routing and terminating points. It is necessary to take these facts into account in engineering toll plant additions so that they will lead toward an advantageous layout for future nationwide dialing as well as meet the needs of the more immediate future. Fortunately, the effect is in the direction of greater concentration of circuits in main routes so that with the new cable and radio facilities available, over-all economy and better service should result.

TYPES OF TRANSMISSION FACILITIES USED AND INCLUDED IN SWITCHING PLAN

The domestic toll network is an outgrowth of the demands of the business and the advance in communication technique over many years. At present, about 100,000 intertoll trunks over twenty-five miles in length and many thousand shorter toll trunks are in service throughout the country. They are provided generally by voice frequency or carrier frequency facilities. The choice of transmission facility on a given route is dependent on a number of factors, such as cost, length of haul, number of trunks in the cross-section, numbers of trunks to be terminated at intermediate points, the types of terrain to be transversed, storm and

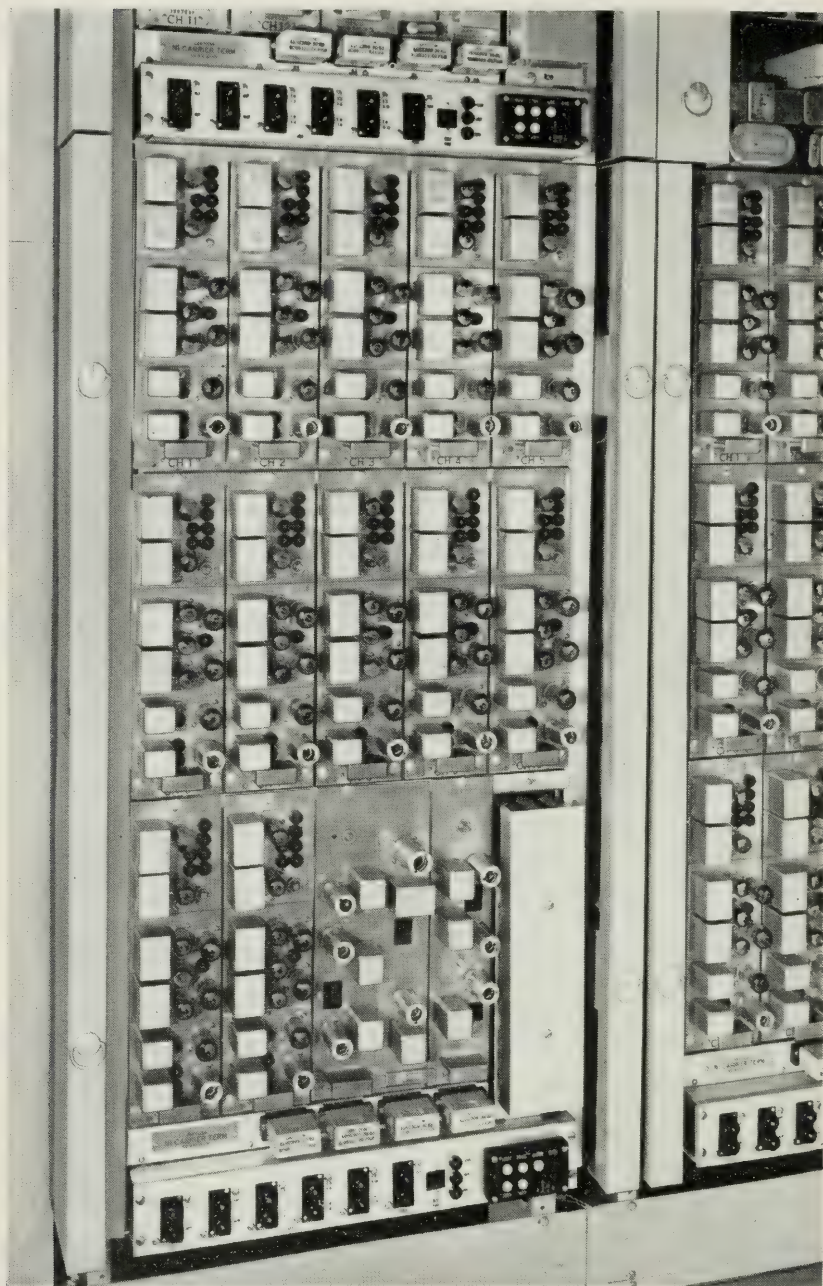


Fig. 8—Terminal equipment of type-N1 cable carrier system. Provides twelve message channels with self contained signaling equipment over two pairs of cable conductors in same sheath.

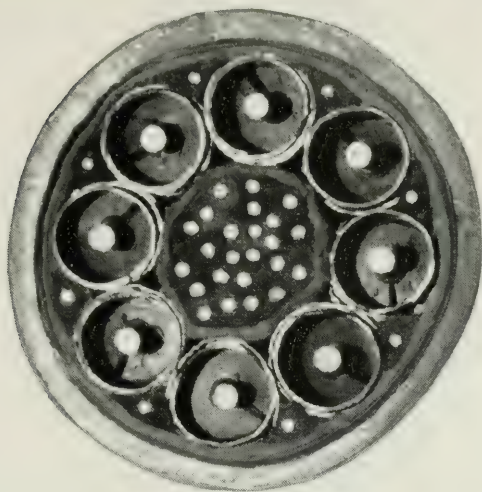


Fig. 9—Coaxial Cable. Cross section of cable containing four pairs of coaxials. Each pair can accommodate one two-way coaxial carrier system.

other conditions affecting service continuity and the transmission requirements of the circuits to be provided.

Voice frequency facilities equipped with repeaters as required are used on both open wire lines and cables. At voice frequencies it is customary to derive three trunks known as a phantom group, from two pairs of open wires or from one "quad" (two pairs) of loaded cable conductors. In general the use of voice frequency facilities is now limited to shorter circuits.

Considerations of economy and service improvement led to the introduction of carrier operation into all types of toll plant as rapidly as the state of the art permitted. This directly affects the toll switching plan from the standpoint of routing and location of switching centers.

At present, carrier systems use four broad categories of facilities: open wire, conventional paired or quadded cables, coaxial cable and radio.

Several types of open wire carrier systems permitting from one to fifteen telephone channels above the frequency band of the voice channel are now in use. In general these systems are used where trunk cross-sections are relatively small and where the terrain and weather conditions make open wire lines economical.

Cable carrier systems at present permit the operation of up to twelve telephone channels on two pairs of cable conductors. These conductors may be in one cable or divided between two separate cables, depending



FIG. 10—Microwave radio relay tower at Cotocin Mountain, Maryland, on a New York-Washington radio route. There are 300 message circuits in service with more planned.

on the type of carrier system (Fig. 8). Coaxial cable transmission systems currently provide up to 600 telephone channels per pair of coaxials (Fig. 9). A new coaxial system, under development, is expected to produce about 1,800 telephone channels per pair of coaxials.

Most of the applications of radio for toll telephone service now contemplated, involve the use of point-to-point microwave systems. By employ-

ing channeling equipment at the terminals of these systems similar to that used for the present coaxial system, each pair of radio channels may provide up to 600 telephone channels. Several pairs of such radio channels may be operated through the same antennas (Fig. 10).

Radio systems are also useful in some cases where the number of toll trunks required is moderate, where diversity is desired or where water or other natural barriers make the provision of wire circuits difficult or impracticable.

The type of facility to be used on a particular route is sometimes affected by requirements for other services such as teletypewriter, television network facilities, program facilities, private lines and other factors.

Trend to Carrier Type Facilities and Advantages to Toll Switching Plan

About 70 per cent of the long haul toll message mileage in Bell Operating Companies is provided on carrier type facilities as contrasted with 7 per cent in 1930 (Fig. 11).

From the transmission standpoint, carrier facilities offer marked ad-

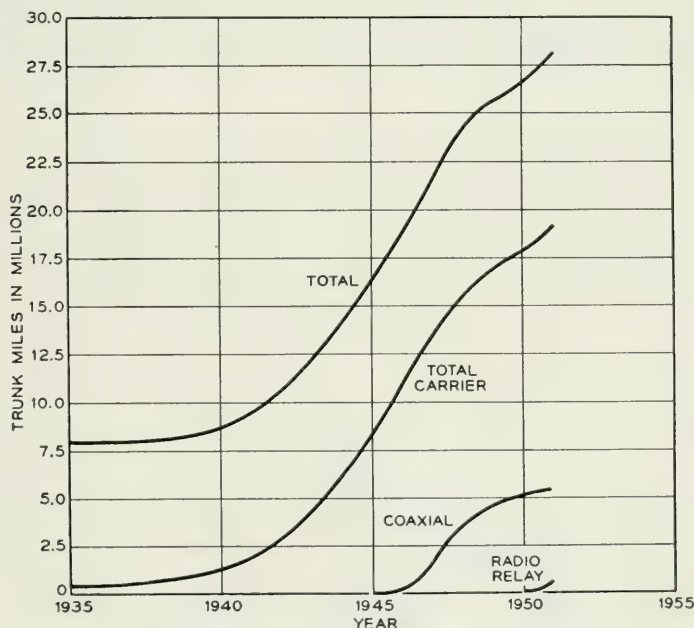


Fig. 11—Growth in Bell System intertoll trunk mileage showing trend toward more extensive use of carrier type facilities.



Fig. 12—Toll switchboard position with key set used for toll dialing.

vantages. They are inherently of the "four-wire" type which minimizes the number of possible singing and echo paths on a circuit. Also, the speeds of propagation over carrier systems are generally higher than over voice frequency systems thereby further minimizing the echo problem. These features are of great advantage in reducing limitations on circuit design and layouts of the general toll switching plan.

Signaling Systems

In addition to the ability to carry messages, intertoll trunks must be provided with suitable signaling facilities.^{6, 7} These must provide a means of: first, attracting the attention of the distant point, either an operator or automatic equipment, to the fact that a connection is to be established; and second, in the case of dial operation, transmitting coded information in the form of pulses for establishing the connection; and third, transmitting a general class of supervisory signals including connect and disconnect signals, on and off switch hook signals, recall signals and

busy signals which are essential to the efficient operation of the switching plant. The circuit design contemplated in the overall plan must take into account this requirement for transmitting signals as well as speech, to obtain accuracy and speed in setting up and taking down connections.

TRANSMISSION DESIGN ASPECTS OF CIRCUITS FOR NATIONWIDE TOLL DIALING

The more extensive use of alternate routing together with the increase in maximum possible number of trunks in tandem associated with nationwide toll dialing, tends to increase the problems of assuring adequate transmission of speech and signals on all possible connections. On the other hand, the use of four-wire switching at important points and the definiteness of the routing patterns permit more effective use of the available facilities and thus tend to simplify the problem. Extensive studies indicate that on the whole, the new toll switching plan will make feasible still further improvements in transmission. This is, of course, a desirable objective.

Transmission Design of Trunks

With dial operation, the number of trunks in tandem in a given toll connection may vary on successive calls. To avoid undesirable transmission contrasts and other adverse effects, it is important that every trunk be designed to operate as closely as possible to the theoretically correct transmission loss. The problem is complicated by the fact that the extent to which the echo, noise and crosstalk will limit the performance of an individual link is not directly proportional to the length of the circuit. In fact, the minimum loss at which a particular circuit used singly or in various built-up combinations can theoretically be operated depends on the number, length and characteristics of the other circuits connected in tandem with it. Arrangements for precisely adjusting the loss in the individual trunks for each call would be complicated. Adequate performance can be achieved however by compromise methods which provide for automatic adjustments in the loss of each trunk in accordance with the following:

1. When a trunk is switched to other intertoll trunks at both ends it is operated at the minimum loss practicable. This loss is known as "via net loss." (VNL)
2. When the trunk is switched to another intertoll trunk at one end only, the loss is increased two db.
3. When the trunk is not switched to another intertoll trunk at either

end a further loss of two db is added. This loss which is four db greater than the via net loss is known as "terminal net loss." (TNL)

The data and methods used in the derivation of the via net loss are rather complex and not within the scope of this paper.

Assignment of Facilities Among Trunks

The definite routing patterns established for the toll machine switching operation impose more severe transmission conditions on certain classes of circuits than on others. For example, a trunk in a "final" group between a TC and a PO can become involved in an eight-link connection, whereas a trunk in a "high usage" group, say, between a PO and another PO will not be involved in more than a three-link connection.

This creates a need and provides an opportunity for allocation of the available facilities among the various trunk groups in a way that will provide the best overall service. For example, to the extent practicable it is desirable to assign carrier grade facilities to trunks in "final" groups that may be involved in connections with the maximum number of links. Facilities with less favorable transmission characteristics may then be reserved for trunks in groups that are used for connections involving fewer links.

TRANSMISSION PERFORMANCE

Table I shows the approximate range of transmission losses between toll centers under the manual plan compared to ranges that appear practicable under the proposed fundamental plan, which, of course, permits more links in tandem.

Trunk Transmission Stability

It is as important that the transmission loss of a trunk used in the contemplated toll dialing network be maintained at or close to its assigned value at all times as that the assigned value be right. On multi-switched connections even a relatively small consistent excess or deficiency in the loss in the individual trunks can accumulate to overall excesses or deficiencies in loss large enough to cause difficulty – by making it hard for people to hear if the attenuation becomes too great or by creating excessive echo, crosstalk or noise if the loss becomes appreciably less than normal.

This subject has been extensively studied for the past several years and it appears that some changes in practices and the introduction of

TABLE I—APPROXIMATE RANGE OF LOSSES BETWEEN TOLL CENTERS IN DB

No. of Links in Intertoll Connection	Manual Plan	Proposed Plan
1	4-12	4-8
2	8-14	5-12
5	9-20	6-13
8	—	7-13

new methods of measuring results will lead to marked improvements. It is of some interest that one of the major factors in securing improvement appears to be the application of a statistical method of evaluating performance along somewhat the same lines as the "quality control" methods used in other fields of industry.

Since, with operator toll dialing only one operator is involved in many connections and with customer toll dialing there is no operator on the connection it is extremely important that everything be right. This is typical of the requirements of any large scale "push button" operation (Fig. 12).

CONCLUSION

The fundamental plans proposed for Telephone Toll Switching provide a basis for the progressive mechanization of toll service. The installation of suitable switching mechanisms at Control Switching Points and the provision of toll trunks utilizing the new instrumentalities will implement the toll switching plan. The plan is sufficiently flexible to adjust for changes in the telephone art as they develop. Also, the plan can fit in with the requirements of those Companies whose plants connect with the Bell operating network should they desire to arrange for operator or customer toll dialing.⁸

Average speed of service will be improved. The flexibility in plant design inherent in the new toll switching plan will increase service security and improve the utilization of the entire toll plant. In addition, adequate provision is made for the progressive introduction of customer toll dialing as this becomes practicable and desirable.

BIBLIOGRAPHY

1. H. S. Osborne, "The General Switching Plan for Telephone Toll Service," *A.I.E.E. Transactions*, **49**, pp. 1549-1557, 1930.
2. C. M. Mapes, "Carrier is King," *Bell Tel. Mag.*, **28**, pp. 191-203, Winter 1949-50.

3. J. J. Pilliod and H. L. Ryan, "Operator Toll Dialing," *Bell Tel. Mag.*, **24**, pp. 101-115, Summer 1945.
4. E. W. Baker, "Toll Dialing is Expanding Throughout the Nation," *Bell Tel. Mag.*, **30**, pp. 253-264, Winter 1951-52.
5. F. F. Shipley, "Automatic Toll Switching Systems." Page 860 of this issue.
6. C. A. Dahlbom, A. W. Horton, Jr. and D. L. Moody, "Application of Multi-frequency Pulsing in Switching," *A.I.E.E. Transactions*, **68**, pp. 392-396, 1949.
7. N. A. Newell and A. Weaver, "Single-frequency Signaling System for Supervision and Dialing over Long Distance Telephone Trunks," *A.I.E.E. Transactions*, **70**, (7 pages), 1951.
8. Articles prepared by American Telephone and Telegraph Company for information of Dial Interexchange Committee of the United States Independent Telephone Association. Published in *Telephony* on dates indicated.
 - a. Nationwide Operator Toll Dialing, January 12, 1946.
 - b. New Toll Switching Plan for Nationwide Dialing, May 10 and 17, 1947.
 - c. Nationwide Toll Dialing—Use of Tandem CDO's, July 3, 1948.

Nationwide Numbering Plan

By W. H. NUNN

(Manuscript received May 15, 1952)

In telephone language a numbering plan gives each telephone in a city, a town, or a geographical area an identity or designation different from that given any other telephone in the same area. There is a wide variation in the types of numbering arrangements in use today in the Bell System, and this paper gives the reasons for this diversity, and examples of the various numbering plans now in use. With the introduction of modern toll switching facilities and the extension of toll dialing to nationwide scope, it was realized that an improvement in the method of dialing toll calls to distant cities was essential in order to realize the maximum speed and accuracy inherent in toll dialing. A nationwide numbering plan covering the United States and Canada has been designed. Each of the more than 20,000 central offices in the two countries are to be given a distinctive designation which identifies that particular office. This designation is to consist of a regional or area code and a central office code. The new switching equipment for the key points in the toll network is being designed so that any toll operator, wherever located, will use the same designation or code for reaching a given office. The combination involved in laying out these areas and the composition of the area codes are presented. A total of 152 codes are available of which approximately 90 are assigned to the present numbering plan areas. Ultimately each central office will be given a type of number consisting of an office name and five numerical digits, such as LOcust 4-5678, in which the first two letters of the office name become the two letters of the central office code. The entire program will take a considerable number of years to realize, but is one which must be accomplished in order to achieve the best results in operator toll dialing and the ultimate goal of nationwide customer toll dialing.

In telephone language a numbering plan is exactly what the name implies, a plan or system of giving each telephone in a city, a town or any geographical area an identity or designation which is different from that given every other telephone in this same area. This designation is the

telephone number; it appears in the directory and in most cities on the telephone instrument itself. It is the address of the telephone in the telephone network. Just as it is essential for efficient postal and delivery service to have streets and house numbers clearly marked, it is important for good telephone service that the telephone numbering plan be such that it will be used with convenience and accuracy by the telephone customer.

A telephone number is comprised of two elements, a designation for the central office to which the telephone is connected and a number within the central office which identifies one particular telephone from all others served by that office. If there is only one central office in the city or town, the office designation is frequently omitted. A dial office is designed to serve up to 10,000 numbers with a limitation of four digits. Typical numbers are therefore MAin 2-1234, ADams-2345, 5-6789 and 3456, the office designations being MAin 2, ADams and 5 with the last four digits in all cases representing the number within the central office.

There is a wide variation in the types of numbering arrangements in use today in the Bell System. This diversity arises from the fact that telephone communities vary greatly with respect to the number of telephones served, ranging all the way from New York City with its more than three million telephones and three hundred central offices to small villages and rural communities with perhaps a few score or a few hundred telephones.

In the 1920's when the Bell System embarked upon its program of converting local offices to dial operation each exchange or city was in general an entity unto itself. Customers dialed local calls within their own city but all calls involving a toll or multi-unit charge required handling by operators for timing and ticketing. There was no advantage, therefore, in making a numbering plan for a given city more comprehensive than required to serve the telephones and central offices in that city with a suitable allowance for the expected growth. Thus there were formed a multitude of local dial communities, large and small, within which customers could dial their own calls and connections between these telephone communities were established by operators.

Over the years these basic numbering plans which were originally established for local dialing have in many of the cities proved inadequate to furnish as many office codes as later events have shown are required. This is due to a variety of causes. The station growth in many places has outstripped all expectations and the number of central offices required to serve this unprecedented demand for service consume many more office codes than the original plans provided for.

In many places local service areas were changed so that customers could call into contiguous exchanges at local rates. To enable customers to dial into these nearby places the original numbering plans required expansion to include this increased number of offices. In addition, with the advance in the telephone art many cities introduced equipment for automatic charging on multi-unit and short haul toll calls so that customers could dial such calls directly instead of placing them with an operator for completion. In order to enable customers to dial these calls, it was necessary to expand the original city numbering plans to encompass wider and wider geographical areas.

In expanding the various types of numbering plans to serve a larger number of central offices than were originally anticipated, various expedients were resorted to. In the largest cities having three-letter office codes a numeral was substituted for the third letter thus very materially increasing the code capacity from about 325 to about 500 and making it possible to form a number of codes using the same office name. The name CANal for example, instead of serving but one office may serve a number of offices, CANal 2, CANal 3, CANal 4, etc. In the medium size cities having two-letter codes, expansion meant adding a digit to the code to all or in some cases to only a part of the offices in the city.

The five-digit places were usually expanded by adding a digit to some of the numbers so that some of the telephones had five digits and others six digits in their numbers.

As a result of choosing originally a numbering plan which at the time seemed adequate and most suitable for the city involved and in many cases being forced to expand to meet changing needs, we now have in the Bell System a considerable variety of different numbering plans. These are given in Table I. The numbering plans given are all adequate to serve the present local dialing needs for the cities in which they appear.

Having reviewed the numbering plan situation as it exists today in the various cities and towns, let us turn to the problem of handling toll calls. Under ringdown operation there is an operator at the outward toll center where the call originates and another operator at the terminating or inward toll center. On built-up toll connections there are additional operators at each intermediate toll switching point. The inward toll operators, who are familiar with the numbering plans in the offices served by their particular toll center, can be relied upon to connect to the desired station even though there is uncertainty on the part of the calling customer or the outward toll operator regarding the precise pronunciation or spelling of the name of the called office or the particular form of numbering system used at the called city.

Under operator toll dialing the inward operator is replaced by dial switching equipment under the control of the outward operator; hence the outward operator has no one to rely upon but herself in completing a toll connection to a distant city. With the present method the operator dials a code for each circuit group in the connection followed by the number of the called party which may consist of any number of digits from three to seven. The operator must refer to her position bulletin or to a routing operator for the correct circuit group codes unless she happens to remember them. Where the office to be reached has central office names, the operator must rely on routing information to determine how many letters of the name are to be dialed. The great variation in the number of digits to be dialed on different calls is a source of some difficulty and confusion to the operators.

The present system of operator toll dialing by which operators use codes depending upon the routes to reach a desired destination, is a great improvement over the old manual handling methods. However, with the introduction of more modern toll switching facilities and the nationwide extension of toll dialing, it was realized that an improvement in the methods for dialing toll calls to distant cities was essential in order to realize the maximum speed and accuracy inherent in toll dialing.

These handicaps in the present toll dialing methods are to be overcome by establishing a nationwide numbering plan covering the United States and Canada by which each of the more than 20,000 central offices in the two countries is to be given a distinctive designation which identifies that particular office and that office only. This designation is to consist of

TABLE I—DIFFERENT TYPES OF NUMBERING PLANS

Place	Directory Listing	Customer Dials	Ordinarily Referred to as
Philadelphia, Pa.	LOcust 4-5678	LO 4-5678	Two-five
Los Angeles, Cal.	PARKway 2345 and *REpublic 2-3456	PA 2345 and RE 2-3456	Combined two-four and two-five
Indianapolis, Ind.	MARKet 6789	MA 6789	Two-four
El Paso, Texas	PRospect 2-3456 and 5-5678	PR 2-3456 and 5-5678	Combined two-five and five digit
San Diego, Cal.	Franklin 9-2345 Franklin 6789	F 9-2345 F 6789	One letter, four and five digit
Des Moines, Iowa	4-1234 and 62-2345	4-1234 and 62- 2345	Combined five and six digit
Binghamton, N. Y.	2-5678	2-5678	Five digit
Manchester, Conn.	5678 and 2-2345	5678 and 2-2345	Combined four and five digit
Winchester, Va.	3456	3456	Four digit
Ayer, Mass.	629 and 2345	629 and 2345	Combined three and four digit
Jamesport, N. Y.	325	325	Three digit

two elements, a regional or area code and a central office code. Any outward toll operator, wherever located, will use that same designation in reaching that office through the dial toll switching network.

In a sense, all of the thousands of offices involved are to be treated as though they were contained in one huge multi-office city. Toll operators will use the area code and the office code in reaching an office situated outside her own numbering plan area, while on calls to points within her own numbering plan area she will dial only the number as listed for toll in the directory. In principle the method employed is to divide the two countries geographically into numbering plan areas and to give each of these areas a distinctive code. Refer to Fig. 1. Within each of these numbering plan areas each office will have a code unlike that of any other office in the same numbering plan area and also unlike any area code. Hence for toll dialing purposes each office will have an area code and central office code which will form a combination unlike that of any other central office in the two countries.

In this geographical division into numbering plan areas, border lines between states and between Canadian provinces have generally been used as numbering area boundaries. Since about 500 central offices are the maximum number which can be served in a numbering plan area, it is necessary to divide the larger and more populous states and provinces into two or more areas making, of course, due allowance for growth. New York state with the largest number of central offices is divided into six numbering plan areas; Pennsylvania, Illinois, Texas and California have four areas each. Other divided states have three or two areas depending upon the number of offices to be served. Approximately 90 areas are being provided, with 14 states and two provinces served by two or more numbering plan areas, the remaining states and provinces by one area each.

In fixing the intrastate numbering plan area boundaries of subdivided states, among other considerations effort was made to avoid cutting across heavy toll traffic routes in order to have as much of the toll traffic as possible terminating in the area in which it originated. The advantage of arranging the numbering plan areas in this manner is readily apparent since on this traffic which does not pass an area boundary the area code is not required.

Let us now consider the composition of the area codes. As indicated previously they must be of a type which will enable the switching equipment to distinguish them from the codes of central offices.

On the telephone dial plate letters are assigned only to the dial positions 2 to 9, inclusive (on some dial plates a Z appears on the 0 position

but the Z is never used in a central office code), hence any office code will always avoid a 1 or a 0 in the first two places. The digits 1 and 0 can therefore be used in area codes to distinguish these from office codes. It is not practical to use them as initial digits of area codes since customers dial 0 to reach operators and the local dial equipment is arranged to ignore an initial 1 for technical reasons. A 1 or 0 in the second place, however, can be employed in an area code without conflicting with any central office codes or interfering with any existing practices. Accordingly the area codes will consist of three digits with either a 1 or a 0 as the middle digit, 516, 201, etc. A few codes of this type are now in use, leaving a practical total of 152 of these area codes available as compared to approximately 90 assigned to our present numbering plan areas. This will provide a comfortable spare for additional future numbering plan areas or possibly for reaching overseas points which may later be incorporated into the toll dialing network.

As shown in Fig. 1, states and provinces such as Montana or Alberta which are contained in a single numbering plan area will have area codes with a 0 as the middle digit to distinguish them from areas in divided states such as Texas where the middle digit will be a 1. This is to enable toll operators to differentiate between the two classifications of areas. On calls to single area states the operators will always know that every call to the state in question uses the one area code, whereas on calls to subdivided states additional information will be required to determine which of the several area codes should be employed to reach the particular destination. It is proposed to show on the operator position bulletin the codes of all single area states and the codes of all frequently called cities in multi-area states. The area codes of the less frequently called places in the multi-area will be obtained from a routing operator.

Within each numbering plan area each of the 500 or fewer offices are to be given a three-digit office code which will be different from that of any other office code in that same area. Ultimately each central office will be given a 2-5 type of number consisting of an office name and five numerical digits, such as LOcust 4-5678, illustrated for Philadelphia. In the larger cities customers will dial seven digits, LO 4-5678, on local calls to numbers in the same exchange. In many of the smaller places the customers on local calls will dial only the numerical digits, the office name being employed for toll dialing purposes only.

Considering the thousands of central offices which now have numbers other than the 2-5 type and the fact that to change existing numbering systems is a difficult and often costly procedure, it will be a number of years before this ultimate objective is realized. As a practical measure,

therefore, it will be necessary during this interim period, before the central office names with the 2-5 type of number are established everywhere, to employ for operator toll dialing office codes which in many cases may not be derived from the customers' telephone number.

In dialing to a combined 2-4 and 2-5 city, for example Los Angeles, the three-digit office code for the Parkway office which has six digits in the local number, will be PAR, whereas to reach the Republic 2 office having seven digits in the local number, the office code will be RE2. To call a telephone in Winchester, Va., with only four digits in the local number, the operator will use a code consisting of numerical digits only, such as 294 which, of course, must be different from every other office code in this numbering plan area. To secure the particular office code to be used in reaching an office where the called number does not furnish complete information, the toll operator must refer to a position bulletin or the route operator. This reference work, of course, takes time and therefore imposes a delay in completing the call.

In addition to giving a distinctive three-digit code to each office within each numbering plan area, each toll center will also be given a three-digit code to enable outward operators to reach inward information, and delayed call operators at toll centers in distant cities. Calls to these operators will be routed in the same manner as calls to customers except that the operator codes will be used instead of a station number and a toll center code in place of a central office code.

The central office names now in use in the various cities in the System were chosen, generally speaking, on the basis of their suitability for customer dialing within the city itself. Many of these names are unfamiliar words to operators and customers in distant cities and the use of these names contributes materially to the operator dialing errors. This situation is gradually being corrected by using for new offices, names from a System approved list and replacing existing names which experience has shown to be particularly troublesome by names from this list.

While numbering plans are important in operator toll dialing, they play an even more essential part in the dialing of toll calls by customers. Operators can be trained to adapt their dialing procedures to the type of local numbering system encountered in the called city even though more time is consumed and more errors result than would be the case if all telephone numbers were of a uniform type. Customers, however, could not be expected to follow any plan which requires a variety of different procedures to be used in reaching different cities. Only a numbering system which is readily understandable and which customers find

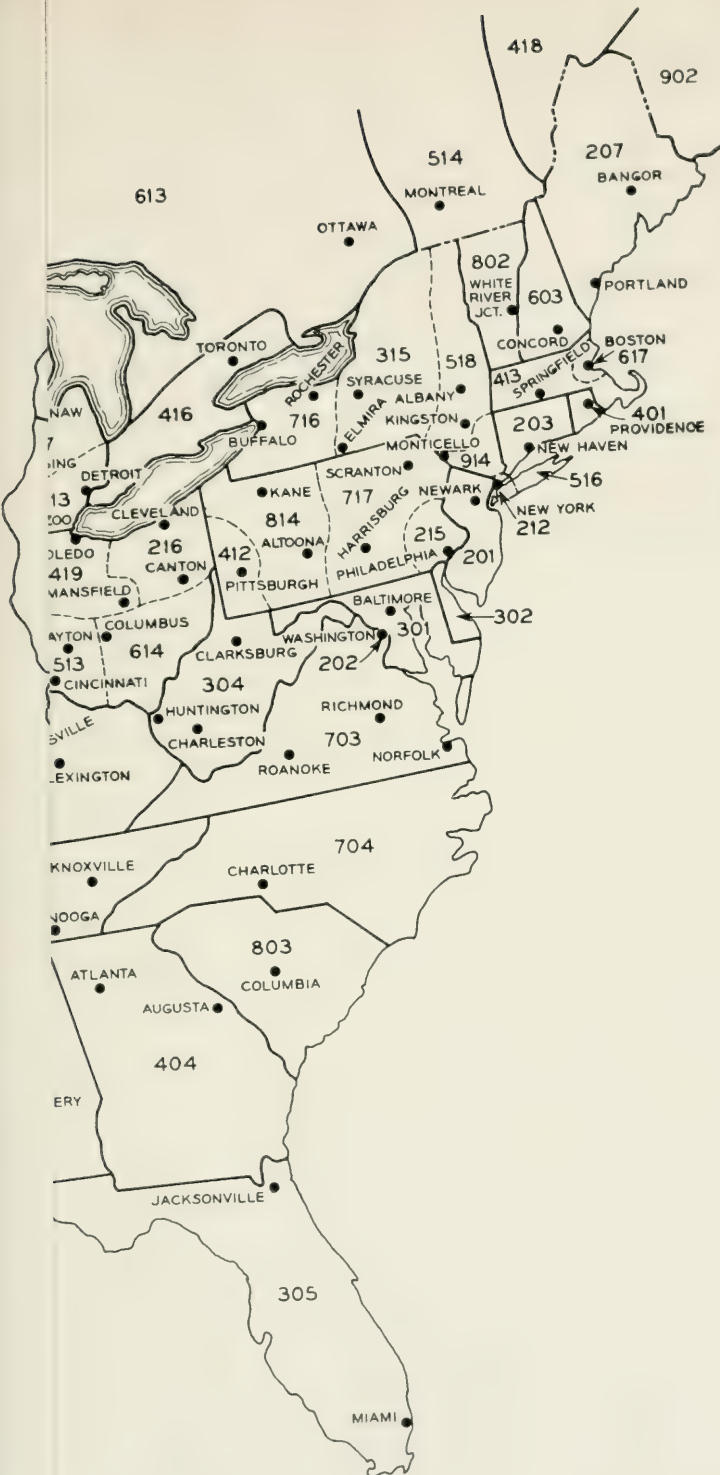
convenient to use and one which they can use with a very high degree of accuracy will suffice. The need for accuracy is readily apparent since with the customer's telephone being given access to the intertoll network without the intervention of an operator, a call which is misdialled can be routed to a telephone thousands of miles from the desired destination.

At present customer dialing of toll and multi-unit calls is for the most part confined to situations where the call can be completed by the use of the number as listed in the directory without any additional digits being dialed. In a few cases as from Camden, N. J. to Philadelphia and certain offices in Northern New Jersey to New York City, the code 11 is prefixed to the listed number. In the case of the current trial of customer toll dialing at Englewood, N. J., the customers are using area codes such as 415 for Oakland, California, 312 for Chicago, etc., dialing only into those cities which now have the 2-5 type of numbering.

From the Englewood experience it can be confidently predicted that this form of dialing, i.e., an area code followed by a telephone number consisting of a uniform number of digits, is one that customers will use with a reasonable degree of convenience and accuracy. The problem therefore to meet the requirements for nationwide customer toll dialing, is to establish universally for all central offices regardless of size and location a uniform pattern of numbering for toll purposes. The only form of number completely filling the needs is the 2-5 system, which is that used in the largest cities today.

Accordingly, in order to implement the program for customer dialing of toll calls on a nationwide basis, it will be necessary to place all telephone numbers on a 2-5 basis with the code of each office different from that of every other office in the same numbering plan area. Thus each of the 50,000,000 telephones in the United States and Canada will have, for toll dialing purposes, a distinct identity consisting of ten digits; a three-digit area code, an office code of two letters of an office name and a numeral, and four digits of the station number within the office. Typical numbers for toll dialing would therefore be 601-CA3-4567 or 317-MA7-6789. As with operator toll dialing, on a toll call which terminates in the same numbering plan area in which it originates, the area code will be omitted and the office code and station number—a total of seven digits will be used.

With this universal 2-5 type of number, local calls in and about the larger and medium sized exchanges will be completed by dialing the entire seven-digit number. For many of the smaller places in the more isolated sections, 5-digit or 4-digit dialing will frequently be employed where this number of digits will be adequate for all of the telephones



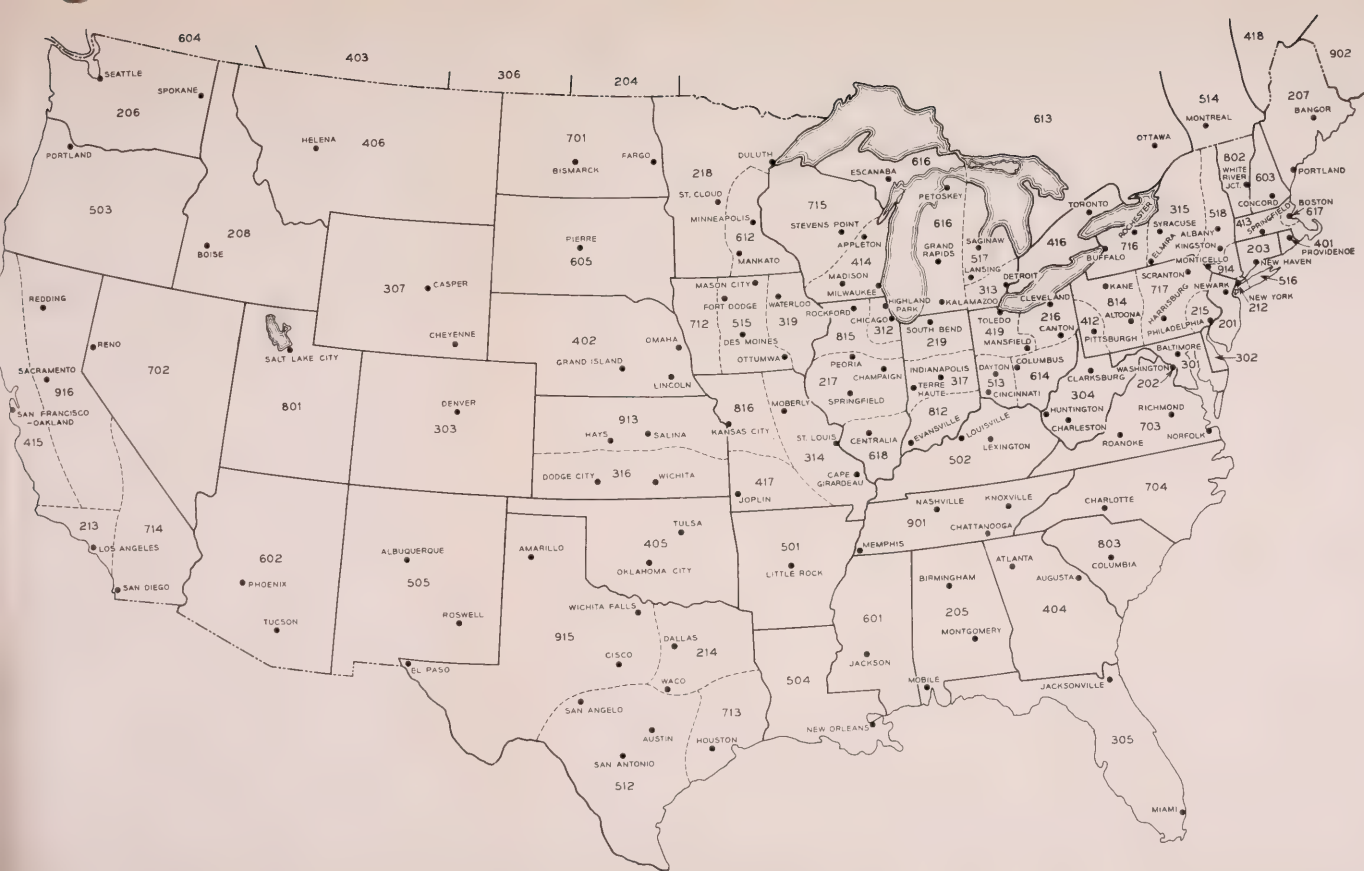


Fig. 1—Nationwide toll dialing areas in the United States and Canada.

in the customers' local dialing area. For these offices with five or four-digit local dialing and for offices in the larger places served by certain types of dial equipment, as they are arranged today, it will be necessary to prefix the dialing of toll calls by a transfer or directing code to permit the customer getting from the local office into the toll network.

Independent of the advantages of a universal 2-5 numbering plan for nationwide operator and customer toll dialing, the Bell System has made considerable progress in this direction over the past several years. New York and Northern New Jersey adopted 2-5 numbering in 1930 in order to take advantage of the flexibility of office code assignments and the large code capacity which this type of local numbering provides. Since World War II many cities and their environs such as Chicago, Boston, Philadelphia, San Francisco, Oakland, Pittsburgh, Milwaukee, Providence and a number of smaller cities have followed suit. Presently about 12 million telephones are in areas which have 2-5 numbering exclusively in addition to perhaps two million telephones with 2-5 numbers in mixed 2-4 and 2-5 areas. Another five million telephones are already planned for conversion to 2-5 numbers within the next several years.

The entire program will take many years to realize but it is one which must be accomplished in order to achieve the best results in operator toll dialing and make it possible for a customer at any telephone in the United States and Canada to reach a telephone anywhere in the two countries by dialing without the assistance of an operator.

in the customers' local dialing area. For these offices with five or four-digit local dialing and for offices in the larger places served by certain types of dial equipment, as they are arranged today, it will be necessary to prefix the dialing of toll calls by a transfer or directing code to permit the customer getting from the local office into the toll network.

Independent of the advantages of a universal 2-5 numbering plan for nationwide operator and customer toll dialing, the Bell System has made considerable progress in this direction over the past several years. New York and Northern New Jersey adopted 2-5 numbering in 1930 in order to take advantage of the flexibility of office code assignments and the large code capacity which this type of local numbering provides. Since World War II many cities and their environs such as Chicago, Boston, Philadelphia, San Francisco, Oakland, Pittsburgh, Milwaukee, Providence and a number of smaller cities have followed suit. Presently about 12 million telephones are in areas which have 2-5 numbering exclusively in addition to perhaps two million telephones with 2-5 numbers in mixed 2-4 and 2-5 areas. Another five million telephones are already planned for conversion to 2-5 numbers within the next several years.

The entire program will take many years to realize but it is one which must be accomplished in order to achieve the best results in operator toll dialing and make it possible for a customer at any telephone in the United States and Canada to reach a telephone anywhere in the two countries by dialing without the assistance of an operator.

Automatic Toll Switching Systems

By F. F. SHIPLEY

(Manuscript received May 12, 1952)

A new automatic toll switching system has been developed by the Bell Telephone Laboratories for use at the most important switching centers for implementing the nationwide dialing program. The job of performing the switching functions at such points is the most comprehensive ever performed by any system, requiring a high order of mechanical intelligence. The new switching system uses crossbar switches for the talking connections and fully exploits the common control principle whereby the equipment used for directing the establishment of connections through the switches is provided in pools common to the office and is used with high efficiency. To perform the complicated translating functions a new device called the card translator has been developed. It uses punched metal cards and an optical system with phototransistors. Routing changes are made by insertion of previously prepared cards in the machine. The switching system was designed with the objective of handling long distance traffic dialed by customers as well as that dialed by operators.

INTRODUCTION

This paper deals primarily with the major switching centers required for the nationwide automatic switching plan. These are called Control Switching Points (CSP's) and are supplied with switching equipment endowed with great versatility and a high order of mechanical intelligence. Mr. Pilliod's paper¹ explains how for purposes of circuit layout and routing, they are assigned different rankings as follows, starting with the lowest ranking: Primary Outlets (PO's), Sectional Centers (SC's), Regional Centers (RC's) and one National Center (NC). Substantially the same equipment is to be provided for all of these centers so that they all will have inherently the same capabilities. They will, however, differ greatly in size. In the United States and Canada, as now envisaged, there will be somewhat under 100 of these CSP's.

The system which Bell Telephone Laboratories developed for use at CSP's and which embodies all of the features required at those important

switching points is based on the Toll Crossbar System² now in service and has been constructed by the addition of the necessary CSP features to the basic structure of that system.

FUNCTIONS OF THE CSP SWITCHING SYSTEM

The system is designed to be suitable for location in either a step-by-step or a panel-crossbar local area. In addition to the functions required for operation as a CSP, it must, of course, perform the normal toll switching functions required of any system for switching the toll traffic characteristic of the locality it serves. These may be stated very briefly.

Ordinary Toll Switching Functions

1. It accepts calls either directly from operators or from senders in distant offices. In the interest of economy it accommodates itself to the signaling language the operator's position or sender is equipped to deliver. Calls from operators may be either in the dial pulse (DP) or multi-frequency³ (MF) form. Calls from senders will be in the MF form.

DP pulsing is the decimal type delivered directly by the dial and is at the rate of about one digit per second. MF pulsing represents a particular digit by a combination of two out of five frequencies in the voice range; it uses one of these frequencies in combination with a sixth frequency to produce a signal indicating the beginning of pulsing, and a different one of the five in combination with the sixth for an end of pulsing signal. It is transmitted from senders at the rate of about seven digits per second. Operators usually key at the rate of up to two per second.

2. The toll switching system completes calls to various types of mechanical toll and local offices and to operators, using the form of signaling dictated by economy for each call. For distant toll offices and local offices using step-by-step equipment DP will be transmitted, for other CSP's and usually for local crossbar offices MF will be transmitted and at manual toll offices an operator will be called in either automatically on seizure of the toll line or by sending a ringing signal over the line, but no pulses will be transmitted. Forms of pulsing different from either of these are used for local panel offices and for local manual offices in panel-crossbar areas.

3. It must transmit signals in one direction for initiating, holding and releasing the connection and in the opposite direction to indicate to the originating end when the called subscriber answers and hangs up. These

signals must be in a form suitable for propagation over the medium which carries them.

4. It must exercise control over the amount of amplification of voice currents introduced at the switching point so that a proper grade of transmission will be furnished.

All of these functions are performed by toll crossbar systems already in service. The features that distinguish the new system are those peculiarly characteristic of CSP operation.

CSP Functions

The following features which will be built into the equipment at Control Switching Points are commonly referred to as CSP features:

1. Storing and sending forward digits as needed.
2. Automatic alternate routing.
3. Code conversion.
4. Six-digit translation.

The first of these features is basically essential for implementation of the plan. The second produces faster service and important economies in outside plant. It also provides protection against complete interruption of service in case of failure of all circuits on particular routes. This aspect of the feature is so important that automatic alternate routing may also be considered essential. The other two features are provided for reasons of economy, and produce economies of such magnitude that they are very much worth while.

1. Storing and Sending Forward Digits as Needed

The necessity of providing this feature in CSP switching systems arises from the nature of the numbering and switching plans. The numbering plan⁴ is constructed with the objective of using a minimum number of digits to give each telephone user in the country a distinctive number.

Numbers delivered to the CSP equipment are in the form ABX-XXXX if the called place is in the same numbering area as the CSP. AB represents the first two letters of any office name and X represents any numeral. If the called place is in another numbering area this set of digits will be preceded by X0X or X1X. X0X or X1X is the area code, ABX the local office code, and these are the digits used for routing purposes. Regardless of the number of switches required to complete the call, these two sets of code digits are all that will be supplied. They are universal codes in that they identify specific destinations — any place

in the United States or Canada – and for a particular destination the same set of digits will be used wherever the call may originate. All CSP's must, therefore, be able to advance a call toward the same place when the same set of digits is received.

To make use of destination codes possible, each CSP must store the digits as received and pass along to the next point whatever digits may be required there for advancing the call. If the next point is a CSP not in the home numbering area of the called place, the complete ten-digit number will be sent forward. If it is a CSP in the home numbering area of the called point the area code will be dropped and the remaining seven digits will be sent forward. That CSP may in turn complete to a local office directly, dropping the office code, or through a step-by-step TO (Tandem Outlet) or TC (Ordinary Toll Center), substituting arbitrary digits for the area or office code, thereby exercising the third of the listed CSP features.

2. Automatic Alternate Routing

The system is arranged to offer the maximum number of alternate routes possible under the switching plan. As explained by Mr. Pilliod,¹ a maximum of five alternates will actually be used. This number is possible, of course, only at PO's since higher ranking CSP's have fewer CSP's above them in the final chain.

3. Code Conversion

This refers to the ability to substitute one, two or three arbitrary digits for the area code, the office code or both. It is economically important to be able to do this because it makes it possible to work with the step-by-step equipment extensively used in local offices and in toll offices in TC's or TO's without the changes in local numbering plans or rearrangements – and in some cases extra selectors – required for the step-by-step TC's or TO's to use ABX codes for routing purposes. Even though eventually all customers are listed as ABX-XXXX and TC's are arranged to use the listed number for routing the calls, this will not be accomplished for some time. Moreover, after such arrangements are in effect there will still be need for code conversion, particularly for routing calls through TO's. Many combinations of digit dropping and substitution are required to cover all possible cases.

4. Six-Digit Translation

When a CSP receives a ten-digit number it is sometimes sufficient to translate only the area code digits and sometimes necessary to trans-

late both the area and office codes. If all points in the called area are reached by the same route out of the particular CSP concerned the area code will suffice for selection of the route. If some points are reached by one route and others by one or more different routes the office code must also be translated to determine which route should be selected.

BASIC ARRANGEMENT

In the CSP switching equipment talking connections are established through crossbar switches.⁵ Incoming and outgoing toll lines and toll connecting trunks are terminated on crossbar switch frames with linkage between them to provide full access. The switches are controlled by equipment common to the office, each item of which is held only long enough to perform its task in setting up the connection.

The major items of common control equipment are senders, markers, decoders and translators. The basic functions of the senders are the same as in other common control systems, i.e., registering incoming digits and sending them out as directed. A departure from prior practice is made in the design of the marker. In other crossbar systems the marker is the principal seat of the mechanical brains. It not only controls the actual establishment of the connection but also does the translating to determine what connection should be established and what information should be passed to the sender for further disposition of the call. In this system the marker still controls the actual setting up of the connection, but it acts on instructions received from the decoder where the major portion of built-in intelligence resides.

The decoder accepts code digits from the sender, translates them, makes selection of alternate routes and gives instructions to markers and senders to enable them to carry out their assignments. To do the translating job the decoder has one, and in some cases two translators permanently associated with it and in addition has access to a common group of translators called foreign area translators which can be used by all decoders as required.

The relationship of the principal elements of the system to each other is depicted in the schematic diagram, Fig. 1.

METHODS OF OPERATION

The manner in which the various elements of the CSP system and the CSP systems at various locations cooperate to implement the nationwide switching plan may best be understood by following the progress of a call which demands the exercise of the characteristic CSP functions.

Assume an outward operator in Atlanta has received a station-to-station call for a subscriber in Monticello, Maine, whose number is ACademy 4-2345, that Monticello is a tributary of Houlton, a step-by-step TC, that Bangor is a step-by-step TO serving Houlton as its home TO and that the circuit groups provided are as indicated in Fig. 2. The dotted lines represent high usage groups and the solid lines final groups.

The Atlanta operator plugs into a tandem trunk to the toll crossbar system in Atlanta, thereby causing a sender to be attached to the trunk through the sender link frame. This causes a lamp signal to be displayed to the operator, indicating that she may key the number. She keys 207-AC4-2345 plus a start signal (signifying end of keying) into the sender and leaves the connection to handle other calls. She will give this call no further attention until the lamp associated with the cord circuit used in establishing it signifies either by flashing that the call was not completed due to a busy condition of the called line or to circuit congestion, or by going dark that the called subscriber has answered and she should start timing the call.

As soon as the area code, 207, is received by the sender it calls for a decoder and gives it the code. The decoder, by means of a self-contained translator finds that the area code is sufficient for routing purposes, that the first choice route is by way of Boston, the second New York and the final St. Louis. Without consulting other circuits it will know in which

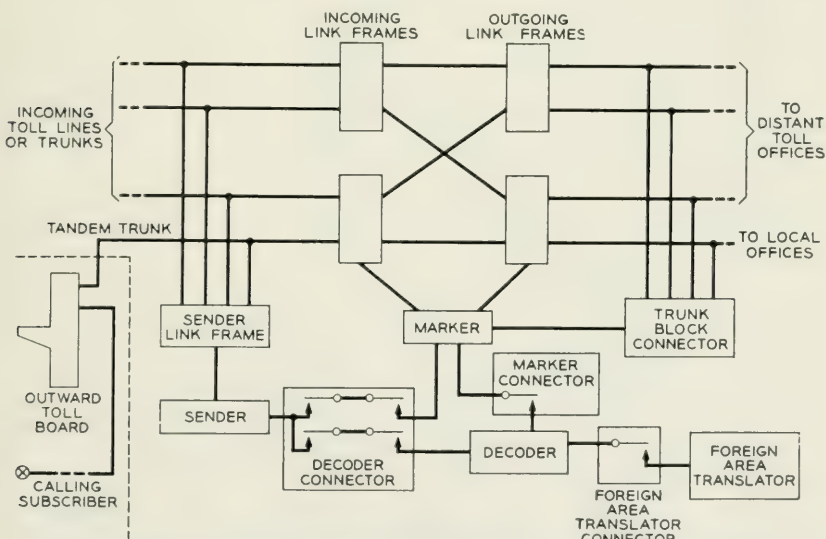


Fig. 1—Schematic diagram of crossbar switching system for CSP's.

of these groups an idle circuit may be found. Let us assume that the circuits to Boston are all busy but there are one or more idle circuits in the New York group. The decoder calls for a marker and tells it which group of leads to test, and also causes the sender to be connected to the particular marker it has selected.

The marker, following instructions from the decoder, is connected to the appropriate trunk block connector. This is one of a group of common circuits giving access to "blocks" of trunks for allowing the marker to locate an idle trunk. The marker examines the test leads of the individual toll lines to New York and as soon as it has selected an idle circuit it so informs the decoder. The decoder then tells the sender to send all

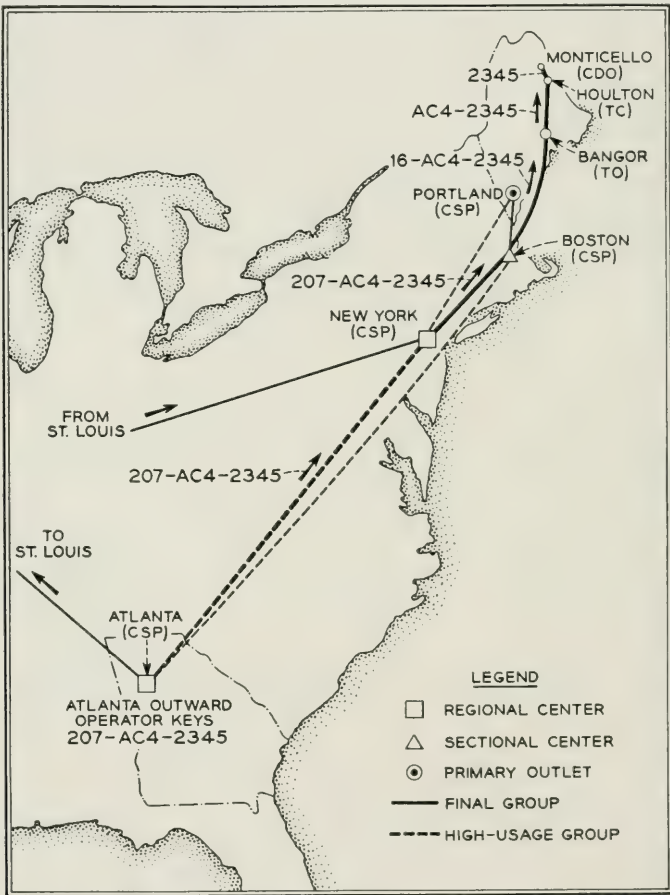


Fig. 2—Call from Atlanta, Georgia to Monticello, Maine.

digits forward by MF and leaves the connection to accept another call. This information is relayed from the decoder to the sender by way of the marker. The work time of the decoder has been in the order of a half second.

The marker determines the identity of the frames on which the incoming and outgoing circuits are located, finds an idle path between the two circuits and sets up the connection. After checking the path through the switches to be sure that there are no troubles it notifies the sender that its task has been completed and then leaves the connection. Its work time has also been in the order of a half second.

In the meantime other digits have been coming in to the sender but it does not wait for all of them to arrive before advancing the call. When the marker selected the circuit to New York a signal was immediately sent forward to summon a sender in the New York switching system. The process of attaching the sender in New York was carried on concurrently with the establishment of the connection through the switches in Atlanta.

When the New York sender is attached a signal is sent to the Atlanta sender to advise it that pulsing may proceed. It immediately sends the area code 207 to New York by MF pulsing and follows it with the remaining digits of the called number, AC4-2345, as they are received from the operator, ending with a start pulse, and then leaves the connection. All common control equipment in Atlanta is now free.

In New York, as soon as the Maine area code is received it is submitted to the decoder. Upon examination of the code the decoder finds that it is insufficient for routing purposes. New York has a direct circuit group to Portland over which traffic to some offices in Maine is routed, but other offices are reached through Bangor by way of Boston. In order to determine which route to take the decoder must know what office is desired. It, therefore, gives the sender a signal saying "come again when you have six digits" and leaves the connection. When the sixth digit arrives the sender again calls for a decoder and gives it the complete code 207-AC4.

The decoder again translates the area code, which now directs it to the foreign area translator which serves the Maine area, and submits the complete code to that translator. From the ensuing translation it learns that the route is by way of Boston and that all digits should be sent forward by MF. It then calls for a marker and releases the foreign area translator.

Subsequent operation is the same as previously described for Atlanta and the complete ten-digit number now arrives at Boston. At that point

both codes are again translated since Boston also has a choice of routes to Maine, and the route to Bangor is selected. The translating equipment in Boston knows that Bangor is in the Maine area and that the area code will, therefore, not be needed. However, since Bangor is a TO having no senders, the Boston sender must pulse forward all of the digits needed to complete the call through switches in Bangor, Houlton and Monticello. It is assumed that Houlton is arranged to route the call to Monticello on receipt of the digits AC4. Numerical digits 2345 will route the call through the Monticello switches to the called customer's line. These digits are all registered in the Boston sender but the digits required to switch the call through Bangor are not and must be supplied. An arbitrary set of digits beginning with "1" can be used for this purpose since no office code begins with "1" and there will, therefore, be no conflict.

The decoder in Boston, therefore, gives the sender the proper set of arbitrary digits, say 16, to be placed ahead of the office code AC4. The sender sends forward by the DP method 16-AC4-2345 driving switches in Bangor, Houlton and Monticello to the called subscriber's line, and ringing starts automatically. The talking connection is now established and the common control equipment at all intermediate points is free.

When the called subscriber answers, the Atlanta operator's cord lamp is extinguished. When he hangs up the lamp lights to denote end of conversation. The removal of the operator's cord automatically releases the entire connection, the release of each link causing the next in line to release.

In setting up this call all of the characteristic CSP features were employed, automatic alternate routing in Atlanta, six-digit translation in New York and Boston, digit storing and variable spilling at all CSP's with substitution of arbitrary digits for the area code at Boston.

TRANSMISSION

All talking connections through the CSP system are made on a four-wire basis, that is, separate pairs of conductors are provided for transmission in the two directions. This is done in order to simplify the problem of maintaining satisfactory balance so that the loss introduced by extra links in a connection can be held to a minimum value. The importance of this feature is emphasized by the fact that the switching plan permits as many as eight intertoll trunks to be connected in tandem for the completion of a call.

The advantages of four-wire switching were fully explained in the paper² on the toll crossbar system now in service.

SIGNALING

In following the progress of the call from Atlanta to Monticello, Maine, it was observed that besides the transmission of information in the form of digits it was necessary to pass a number of control and supervisory signals over the toll lines. These included seizure and disconnect signals in the forward direction and switchhook supervisory signals and sender attached signals in the reverse direction. On some calls it is also necessary to send flashing signals to indicate busy lines or trunks and ringing signals in either direction when operators are called in at intermediate or terminating points to assist in establishing connections.

For the early toll dialing installations the signaling method most widely used was the composite method whereby signaling channels for the three circuits of a phantom group are derived from three of the conductors with the fourth being used for earth potential compensation. Direct current is used for signaling. This is a simple, reliable and economical method of signaling and will continue to be used on circuits where it can be applied.

Where circuits are obtained from carrier systems, however, conductors are not available in sufficient numbers for signaling channels and other methods must be employed. Since carrier is used almost exclusively on the long haul circuits it was necessary to provide a signaling system to accompany it before toll dialing could be expanded beyond networks of limited range. To meet this situation a system⁶ using a frequency of 1600 cycles was developed and has been in service since 1948. Signaling is done by application and removal of the 1600-cycle signaling current. The system is used in the same manner as the composite signaling system, to carry dial pulses as well as supervisory signals when used on circuits that require it. The set of leads brought out of the signaling unit are identical in function to those brought out of the composite signaling unit so that toll line relay circuits will operate in the same manner with either type of signaling.

Since 1600 cycles is in the voice range the signaling current can be carried over the same channel that carries the speech current but the signaling circuits must, of course, be protected against false operation due to speech and precautions must likewise be taken to insure that the signaling tone does not interfere with speech. Protection against interference between signaling and speech is more difficult at 1600 cycles than at higher frequencies because there is more energy in voice currents at the lower range. That value was chosen, nevertheless, so that it would be possible to operate over the narrow band circuits that were established to relieve shortages occasioned by the war.

A new 2600-cycle system to be used only on the broader band circuits has since been developed. It is simpler and more economical than the 1600-cycle system. The older carrier systems, having been designed when practically all toll operation was by the ringdown method, made no provision for signaling since that was all done by short applications of the 1000 cycles when there was no speech on the line. Some of the new carrier systems for short haul applications are designed to provide their own signaling channel for each voice channel.

PRINCIPAL ELEMENTS OF THE CSP SYSTEM

1. Crossbar Switch Frames

Crossbar switches are used for incoming and outgoing link frames on which the trunks (both toll lines and trunks to and from local offices and switchboards) are terminated, and for sender link frames used to give trunks access to senders. These frames are similar to those in the toll crossbar systems now in service. Since they have been described in a previous paper¹ they will be passed over with only a mention of their capacity.

Each incoming or outgoing link frame normally has terminals for 300 trunks. As many frames are provided as required for the size of the office. In the smaller offices one train of switches with complete interconnection of incoming and outgoing frames is provided. In the larger offices two trains each with its own set of markers are provided. When this is done the incoming trunks are multiplied to both trains and an extra build out bay is provided on the incoming frame to provide 400 terminals per frame. Since each train has a theoretical limit of 40 incoming and 40 outgoing frames the maximum size of an office is theoretically 80 of each. Practical considerations, however, such as the number of markers that can be efficiently operated in a group and the maximum size office it is feasible to operate as a single administrative unit will limit an installation to about 60 incoming and 60 outgoing frames.

The sender link frame gives 100 trunks access to 40 senders.

2. Senders

Two separate groups of incoming senders are provided, one to receive DP and the other MF pulsing. Whether the system is installed in a step-by-step or a panel-crossbar area both groups of senders will always be needed. MF will be received from senders in other CSP's and from switchboard positions. DP will be received from switchboard positions

at TC's not equipped to send MF and in some cases from dialing A boards in the local area of the CSP itself.

Aside from the type of pulses received the functions of the two senders are identical. They have a capacity for receiving and sending eleven digits. They must register arbitrary digits given them by the decoder and send them out as directed. They will send out digits by either the DP or the MF method as required to control switches in distant offices, and in some installations will also send digits to an outgoing sender in the same office by the dc key pulsing method, which employs direct current in various combinations of value and polarity through a pair of conductors.

When the CSP is in a panel-crossbar area a group of outgoing senders is provided to transmit either the type of pulses required by the equipment in local panel offices or the type used to reach manual offices.

3. Markers

The marker has been stripped of its usual translating functions and performs most of its duties on instructions from the decoder. It is told what leads to test for idle circuits and where they are to be found in the trunk block connector, but having found an idle circuit it carries on the process of setting up the connection independently of the decoder. Having contact with both the incoming and outgoing trunks through connecting circuits, it determines what frames they are located on, connects itself to those frames, selects a path through them and sets up the connection.

In a single-train office one group of markers common to the office is provided. In a two-train office there is a group of markers associated with each train of switches.

4. Decoders

A single group of decoders serves the entire office whether one or two trains of switches are provided. An important element of the decoder is the translator which will be discussed separately.

The decoder contains several hundred relays. A large group is used for registering the information furnished by the translator. Others use this information to control the action of the markers and senders.

One group of decoder relays which is of particular interest is the array used for automatic selection of alternate routes. It is composed chiefly of one relay for each CSP to which the office has a direct group of toll lines. The relays are arranged in an orderly pattern simulating the

pattern of the CSP network for the country as seen from the CSP concerned and are interconnected in a pattern of progression corresponding to the fixed order of alternate route selection. Group busy leads from the toll line groups are connected to the contacts of the relays in such a manner that if a group is busy the relay corresponding to the next choice route in the chain will be operated. In this way the lowest choice route having an idle circuit will be speedily selected without testing individual trunks of separate groups. The decoder learns from the translator which relay in the array to operate first and the choice of the best route available follows automatically. The principle will be readily understood by reference to the simplified sketch in Fig. 3. Contacts not shown on the relays cause the translator to select the route corresponding to the last relay operated in the chain.

5. Translators

The magnitude of the translating job for nationwide dialing led to the decision to develop a new translator operating on a principle radically different from that employed in other crossbar systems. In previous systems translation is done by relays. The code digits – never more than three – operate a group of relays which cause a single terminal corresponding to the code to be selected. A cross-connection is made between

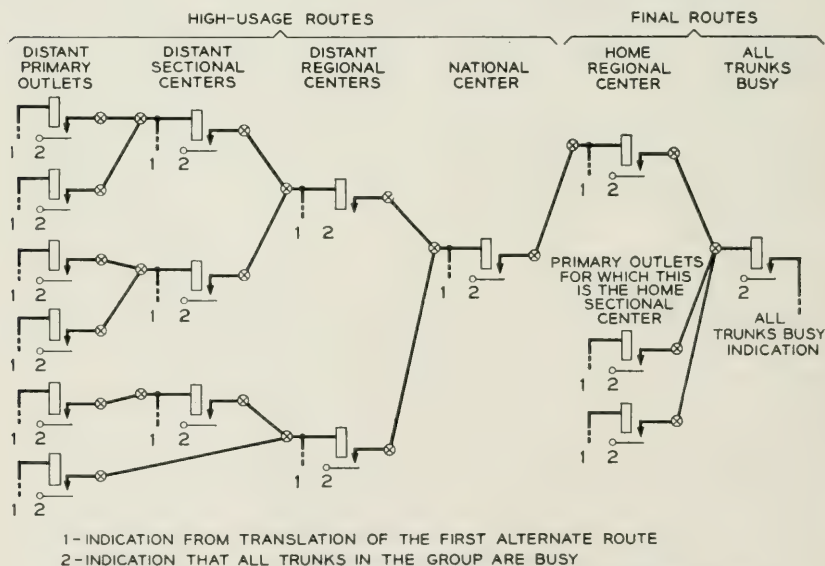


Fig. 3—Alternate route array for the decoder at a sectional center.

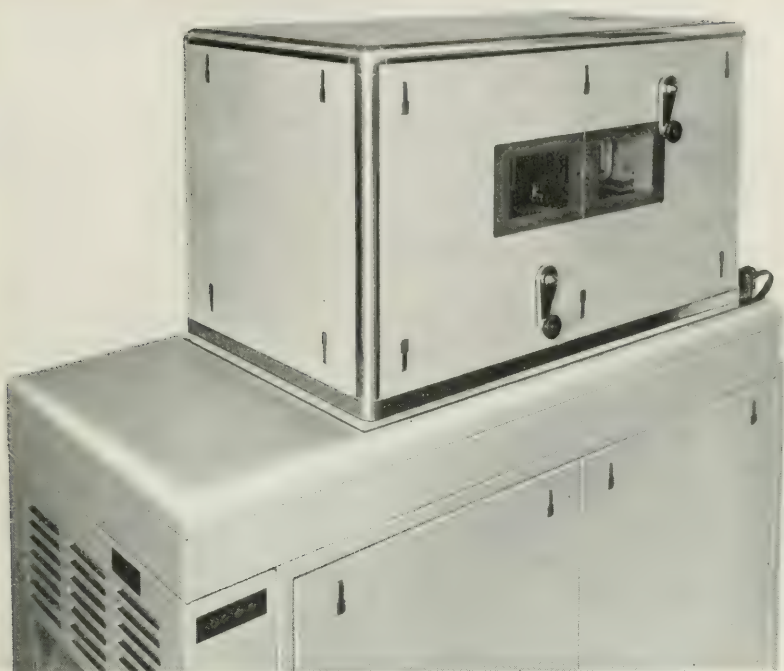


Fig. 4—Card translator.

the code point and a route relay associated with the trunk group to be selected. The route relay has a number of contacts which are cross-connected to supply the information required for proper routing of the call. When changes in routing or equipment location of trunks within the office are made it is necessary to change cross-connections.

With the nationwide dialing plan in operation routing changes or opening of new offices in one part of the country will necessitate translator changes in many offices, some of them far removed from the scene of the event that forces them to be made. The changes in any one CSP will, therefore, be frequent and to make them by running cross-connections would be cumbersome and expensive. The new translator uses punched cards instead of relays, making it possible to effect changes by the simple process of removing old cards and inserting new ones in a machine. This can be done in a very short time and not only saves labor but requires less out-of-service time for the equipment. Fig. 4 is a photograph of the machine.

A metal card about 5 by $10\frac{3}{4}$ inches is provided for each area code and also one for each office code that must be translated in a particu-

lar CSP, the cards representing destinations. The capacity of a single machine is about 1000 cards. The cards are lined up in a box as in a filing drawer, with tabs along the bottom of the card resting on select bars which run the length of the box. One-hundred and eighteen holes are punched out in all cards in fixed positions so that in the normal condition 118 tunnels are formed from one end to the other. A light source at one end of the box shines through the tunnels upon phototransistors (Fig. 5) at the other end but the phototransistors are disabled until, concurrently with the dropping of a card, voltage is applied to them.

All tabs along the bottom of the card are cut off except those which serve to identify the particular card. When a code is presented to the machine a combination of select bars corresponding to the code is

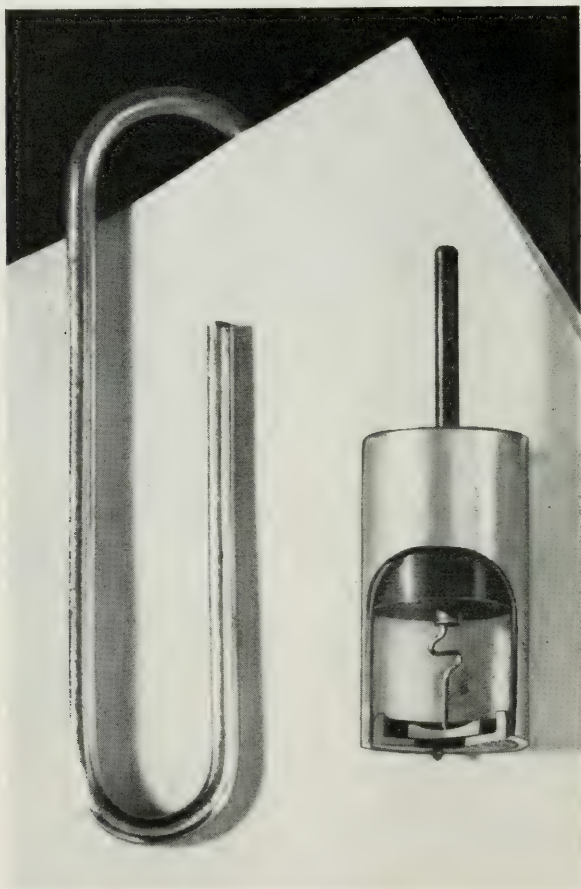


Fig. 5—Transistor.

lowered. The card having all tabs cut off except those resting on the lowered bars will drop but all other cards will remain in place. If nothing further were done the dropping of the card would cut off all light channels but on each card some holes are enlarged and through these holes the light continues to shine, energizing the corresponding phototransistors. The combination of enlarged holes furnishes all of the information needed for routing the call to the destination represented by the card.

Fig. 6 shows the functions of the various groups of tabs and holes. The designations will not appear on the actual card. Fig. 7 is a photograph of an actual card prepared for use.

a. Selecting Tabs - Input Information. The sole use of the information presented to the card translator is to enable it to select the proper card. The information presented is in the form of code digits with accompanying indications of their nature. The information is recognized by cutting off tabs along the bottom of the card in proper combinations.

The groups of tabs labeled A, B, C, D, E and F are for the six code digits. For each digit used two tabs are left since the digits are registered in the sender on a two-out-of-five basis and the leads from the sender will operate the select bars directly. If the card represents an ordinary three-digit code all tabs will be cut off except two each of the A, B and C tabs, two of the four CG tabs and perhaps either the VO or NVO tab. The CG (card group) tabs are used in combination to indicate three-digit, six-digit and alternate route card groups. The VO and NVO (via only and not via only) tabs are used when the group of toll lines over which the call will be routed is divided into one subgroup of a transmission grade suitable for only terminal traffic and another subgroup for either terminal or switched traffic. If the card represents an ordinary six-digit code two tabs will be left in each of the digit positions, and a different pair in the CG group.

b. Punch Holes - Output Information. The output information from the card translator is recognized in the decoder and marker by relays operated in the combinations set up by enlargement of associated holes in the card. The output from the phototransistors is amplified by other transistors to fire cold cathode tubes which in turn operate the relays.

The pretranslation group on the top line of Fig. 7 indicates how many digits the sender must supply for a complete translation. The term "pretranslation" implies that further translation is required. This is not always true. In many cases only the first three digits need to be translated and all information needed for routing the call is supplied by this card. In many cases the six digits of the area and office code are needed and the routing information will be on another card to be selected

PRETRANSLATION										OGT APPEARANCE				TRAF. SEP. PC				TRK. GRP. PC & OF													
TRANSLATOR BOX NUMBER				AREA CODE CONTROL				ROUTING INSTRUCTIONS				CODE CONVERSION				GROUP START				GROUP END											
NCA	CA4	CA5	CA6	IT	TC	ITC	HB	BT0	BT1	BU0	BU1	BU2	BU3	BU4	BU7	CLT0	CLT1	CLU0	CLU1	CLU2	CLU3	CLU4	CLU7	CDLC	TS0	TS1	TS2	TPC	TPO	TP1	TP2
																		IND1													
																		IND2													
																		IND3													
																		IND4													
																		IND5													
																		IND6													
																		IND7													
																		IND8													
																		IND9													
																		IND10													
																		IND11													
																		IND12													
																		IND13													
																		IND14													
																		IND15													
																		IND16													
																		IND17													
																		IND18													
																		IND19													
																		IND20													
																		IND21													
																		IND22													
																		IND23													
																		IND24													
																		IND25													
																		IND26													
																		IND27													
																		IND28													
																		IND29													
																		IND30													
																		IND31													
																		IND32													
																		IND33													
																		IND34													
																		IND35													
																		IND36													
																		IND37													
																		IND38													
																		IND39													
																		IND40													
																		IND41													
																		IND42													
																		IND43													
																		IND44													
																		IND45													
																		IND46													
																		IND47													
																		IND48													
																		IND49													
																		IND50													
																		IND51													
																		IND52													
																		IND53													
																		IND54													
																		IND55													
																		IND56													
																		IND57													
																		IND58													
																		IND59													
																		IND60													
																		IND61													
																		IND62													
																		IND63													
																		IND64													
																		IND65													
																		IND66													
																		IND67													
																		IND68													
																		IND69													
																		IND70													
																		IND71													
																		IND72													
																		IND73													
																		IND74													
																		IND75													
																		IND76													
																		IND77													
																		IND78													
																		IND79													
																		IND80													
																		IND81													
																		IND82													
																		IND83													
																		IND84													
																		IND85													
																		IND86													
																		IND87													
																		IND88													
																		IND89													
																		IND90													
																		IND91													
																		IND92													
																		IND93													
																		IND94													
																		IND95													
																		IND96													
																		IND97													
																		IND98													
																		IND99													
																		IND100													

Fig. 6—Card layout.

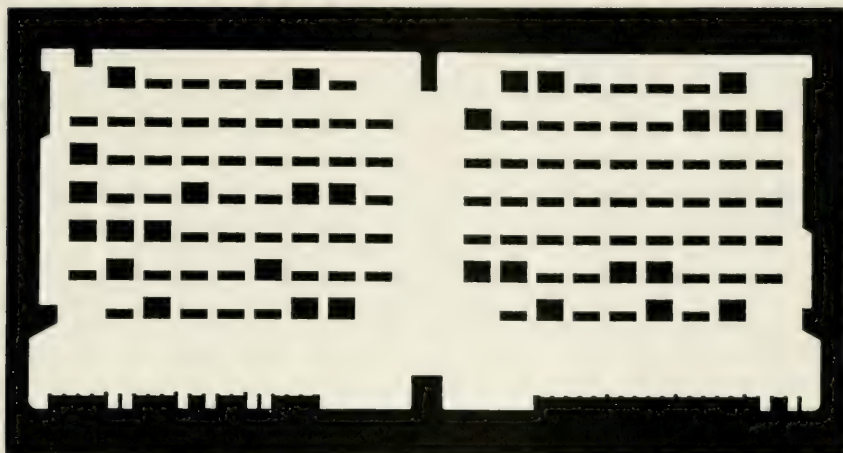


Fig. 7—Punched card.

later. For certain calls such as calls to operators only four or five digits are needed. These are treated as six-digit calls by having the sender supply the extra one or two digits to fill the complement. The NCA hole enlarged means “no come again”, that is, three digits are sufficient, and translation will proceed. The other holes enlarged mean respectively “come again when you have four, five or six digits”, and no further translation is done until the sender comes back a second time, probably to a different decoder, with an indication that it has the required number of digits.

The OGT appearance holes are used in a two-train office to tell which train the outgoing trunk appears on and enable the decoder to select a marker in the proper group.

The remaining holes on the top lines are for controlling operation of traffic meters.

The translator box number holes in the second line are punched on the area code cards to indicate which machine contains the individual cards for the called area when six-digit translation is required.

The IND1 hole in the second line and the IND2 hole in the fourth line are index holes and are never enlarged. Any card that drops will always cut off the light through those channels. This serves as an indication that a card has actually dropped and that the phototransistors associated with the other holes should be prepared for action. The index holes also aid in trouble detection and in proper disposition of calls where cards are deliberately omitted or where operators have dialed a blank code in error.

The class holes indicate such things as type of pulsing and nature of the signaling channel used on the trunk group out of the office.

The area code control holes in the third line are to tell the decoder what to do about dropping or spilling forward an area code registered in the sender or supplying an area code when none is registered. This information is needed primarily in connection with alternate routing.

The alternate route pattern number holes tell the decoder at what point to enter its chain of alternate route relays for the first choice alternate. Provision is made for a maximum of 100 entry points.

The holes on the fourth line are for making proper disposition of calls when no circuits are available on any routes, telling how many digits to expect on certain calls and other items of a detailed technical nature.

The code conversion holes on the fifth line supply the arbitrary digits to replace code digits on calls routed through step-by-step TO's or TC's. Provision is made for one, two or three digits as required.

The variable spill control holes in the sixth line tell whether to spill all digits received, skip the first three code digits or skip six code digits.

The remaining holes define the location on the equipment of the test leads for the trunk group over which the call will be routed.

The notches around the edges are used for proper positioning and removal of cards.

An individual card is removed from the stack by first keying the code to cause it to drop so that it may be identified. Since a card can easily be located in this manner it is unnecessary to keep cards in any ordered position in the box.

At least one translator is provided in every decoder. It contains the cards for all offices in the home numbering area of the CSP, for certain operator codes, the single three-digit card for each toll numbering area and a card for each toll line group out of the office that can be used as an alternate route. If there are other areas to which the volume of traffic is very high and for which six-digit translation is required the cards for those areas are put in a second machine in each decoder. Cards for other areas are put in foreign area translators common to the office and accessible to all decoders on a one-at-a-time basis. An emergency translator is provided to permit removal of all cards to it from any translator which may require prolonged repair work.

6. Traffic Control Panel

The traffic control panel is located in the operating room. The equipment in it consists of a key for each group used as an alternate route. When a particular key is operated no alternate routed traffic will be

offered to the group represented by it nor to any group above it in the fixed alternate route pattern. This is done to relieve offices which are overloaded by either unforeseen or predicted traffic peaks.

MAINTENANCE

The maintenance facilities for the new CSP system are basically similar to those of the older toll crossbar system with the necessary addition of equipment to test the new features introduced. The sender test frame is, of course, obliged to test the CSP features added to the sender and the trouble indicator frame is changed to operate with the new decoders, translators and markers.

In place of the lamp trouble indicator the new trouble recorder introduced with the latest local crossbar system⁷ is used. Whenever trouble is encountered it punches on a card a record of the circuits involved and of the important events that had occurred in the progress of the call, as an aid to the maintenance man in locating the trouble. A sample trouble recorder card is shown in Fig. 8.

Automatic equipment for testing the operation and transmission features of intertoll trunks has also been designed both for the older systems and for the new CSP system.

SWITCHING ASPECTS OF CUSTOMER TOLL DIALING

In the course of developing the switching system for CSP's the requirement for handling long distance traffic dialed by customers as well as that dialed by operators was kept in mind. The trial of long distance customer dialing now in progress in Englewood, N. J., confirms the soundness of the basic plan and exemplifies the principles involved in full realization of the plan. With a toll network laid out to accept a distinctive ten-digit number for any telephone in the country and route it to the proper destination, the remaining tasks to be performed are to provide for delivery of the number to the toll network from the customer's dial instead of from an operator and to provide an automatic record of the call for charging purposes.

In Englewood both tasks were quite easily performed. The Englewood local office equipment is of the most modern type⁷ and includes AMA⁸ facilities. When it was in the development stage the ultimate requirement for nationwide customer dialing was foreseen and provision was made for expanding the digit capacity of the switching equipment at small expense. Also the designs of the accounting center were such that corresponding changes could readily be made. In the new local office switching

system, arrangements were included for sending forward the complete number, as received, to the toll office by MF pulsing. The system was also designed to be capable of automatic alternate routing and this feature is used in the trial.

Expansion of the program will, of course, demand that similar arrangements be provided for the older types of local switching systems already in service. More extensive modification will be required to make them capable of giving the customer the same service. For them, as for the most modern system, however, AMA equipment is admirable for recording the information necessary for charging for the calls.

The requirement for customer toll dialing that senders (or directors) and recording equipment be provided has a bearing on the type of equipment used at TC's and TO's. For calls handled by operators and for calls received by the customers through such offices the only disadvantage of step-by-step equipment without senders at those points is that the CSP equipment at other points must be somewhat more complicated and expensive than it would otherwise need to be. But with customer dialing, if senders and recording equipment are not provided either in the local office or in the TC or TO, the calls must be routed by the most direct means possible to a CSP where such equipment is provided. Thus some advantages that might be gained from having them at the TC or TO would be lost:

1. In some cases an indirect route to the CSP would need to be taken for the sole purpose of recording the call. For example, a call which might normally be switched from a TC through a TO to another TC would need to be connected to the CSP for making the record.

numbering plan covering the entire country with a minimum number of digits to give each customer a distinctive number. It also obviates the need for extra expense to make step-by-step toll offices satisfactory operating elements of the plan in those locations where CSP features are not essential.

The automatic and almost instantaneous selection of alternate routes makes it possible to give virtual no-delay service without greatly increasing the cost of outside plant and to make multi-switch connections at a speed comparable to that for local service.

The translating equipment simplifies administration of the plan which demands coordination of activities on a nationwide basis.

The numbering plan, the switching plan and the CSP equipment which implements them make it feasible to offer nationwide dialing service to customers without the aid of operators when automatic charging facilities and local office switching arrangements for handling the three extra digits of the national number are provided. It will be readily appreciated that so far as the CSP switching equipment is concerned it is immaterial whether the digits it receives come from an operator or from a customer. The call will be routed to its destination and supervision for charging purposes will be furnished in the same manner in either event.

The new system represents an important step in the process of continually improving the long distance switching methods of the Bell System with consequent improvement of the service to all telephone customers in the United States and Canada.

REFERENCES

1. J. J. Pilliod, "Fundamental Plans for Toll Telephone Plant," pp. 832 of this issue.
2. L. G. Abraham, A. J. Busch and F. F. Shipley, "Crossbar Toll Switching System," *A.I.E.E. Transactions*, **63**, June Section, pp. 302-309, 1944.
3. C. A. Dahlbom, A. W. Horton, Jr. and D. L. Moody, "Application of Multi-frequency Pulsing in Switching," *A.I.E.E. Transactions*, **68**, June Section, pp. 505-510, June 1949.
4. W. H. Nunn, "Nationwide Numbering Plan," pp. 851 of this issue.
5. F. J. Scudder and J. N. Reynolds, "Crossbar Dial Telephone Switching System," *A.I.E.E. Transactions*, **58**, May Section, pp. 179-192, 1939.
6. N. A. Newell and A. Weaver, "Single Frequency Signaling for Telephone Trunks," Presented at Winter General Meeting of A.I.E.E., Jan. 31, 1951.
7. F. A. Korn and J. G. Ferguson, "The Number 5 Crossbar Dial Telephone Switching System," *A.I.E.E. Transactions*, **69**, First Section, pp. 233-254, 1950.
8. J. Meszar, "Fundamentals of the Automatic Telephone Message Accounting System," Presented at the Winter General Meeting of A.I.E.E., Jan. 31, 1951.

Mathematical Theory of Laminated Transmission Lines—Part I

By SAMUEL P. MORGAN, JR.

A mathematical analysis is given of the low-loss, broad-band, laminated transmission lines proposed by A. M. Clogston, including both idealized parallel-plane lines and coaxial cables. Part I deals with "Clogston 1" lines, which have laminated conductors with a dielectric, chosen to provide the proper phase velocity for waves on the line, filling the space between the conductors. Part II will treat lines having an arbitrary fraction of their total volume filled with laminations and the rest with dielectric, and will be concerned in particular with "Clogston 2" lines, in which the entire propagation space is occupied by laminated material.

The electromagnetic problem is first formulated in general terms, and then specialized to yield detailed results. The major theoretical questions treated include the determination of the propagation constants and the fields of the principal mode and the higher modes in laminated transmission lines, the choice of optimum proportions for these lines, the calculation of the frequency dependence of attenuation due to the finite thickness of the laminae, the increase in loss caused by improper phase velocity (dielectric mismatch) in Clogston 1 lines and by nonuniformity of the laminated material in Clogston 2 lines, and the effects of dielectric and magnetic dissipation.

TABLE OF CONTENTS

I. Introduction	884
II. Wave Propagation Between Plane and Cylindrical Impedance Sheets.	887
III. Surface Impedance of a Laminated Boundary	896
IV. Principal Mode in Clogston 1 Lines with Infinitesimally Thin Laminae	908
V. Effect of Finite Lamina Thickness. Frequency Dependence of Attenuation in Clogston 1 Lines.	921
VI. Effect of Dielectric Mismatch.	931
VII. Dielectric and Magnetic Losses in Clogston 1 Lines.	940
Appendix I: Bessel Function Expansions.	944
Table of Symbols.	946

I. INTRODUCTION

A recent theoretical paper¹ by A. M. Clogston presents the very interesting discovery that under certain conditions skin effect losses in the conductors of a transmission line at elevated frequencies can be much reduced by laminating the conducting surfaces, parallel to the direction of current flow, with alternate thin layers of conducting and insulating material. The requirements are that the thickness of each conducting layer must be considerably smaller than the skin depth in the conductor, and the phase velocity of waves on the transmission line must be held very close to a certain critical value, which depends on the relative thicknesses and the electrical properties of the conducting and insulating layers. Under these conditions the "effective skin depth" of the laminated surface is greatly increased; in other words, the eddy currents induced by a high-frequency alternating field will penetrate much farther into such a laminated structure than into a solid conductor, with consequent marked reduction of ohmic losses in the metal. The metal losses can also be made to vary much less with frequency, over a fixed band, than the ordinary skin effect losses, which are known to be very nearly proportional to the square root of frequency.

Clogston goes on to show that a laminated material composed of alternate thin conducting and insulating layers may itself be regarded as a transmission medium. For example, if the space in a coaxial cable which is ordinarily occupied by air or other dielectric be filled with a large number of coaxial cylindrical tubes which are alternately conducting and insulating, the cable will propagate various transmission modes, and under the proper circumstances some of these modes will exhibit lower attenuation constants than the transmission mode in a conventional coaxial cable of the same size at the same frequency.

Experimental verification of Clogston's theory of laminated conductors has been obtained² at the Bell Telephone Laboratories, and the transmission properties of a line filled with laminated material have also been measured at these Laboratories and found in reasonable agreement with theory. However experiments with structures as complex as those proposed by Clogston are by no means simple, and the experimental work on laminated conductors is still in an early, exploratory stage. Inasmuch as the experiments are necessarily time-consuming, it has been thought

¹ A. M. Clogston, *Proc. Inst. Radio Engrs.*, **39**, 767 (1951), and *Bell System Tech. J.*, **30**, 491 (1951). References will be to the *Bell System Technical Journal* article, although except for equation numbers the two papers are identical.

² H. S. Black, C. O. Mallinckrodt, and S. P. Morgan, Jr., *Proc. Inst. Radio Engrs.*, **40**, p. 902 (1952).

desirable to carry out simultaneously as complete a theoretical treatment of Clogston-type transmission lines as possible. Clogston's original paper brought out the fundamental ideas by analysis of idealized transmission lines bounded by infinite parallel planes. The present paper considerably extends the theoretical analysis of parallel-plane systems, and also treats laminated transmission lines bounded by coaxial circular cylinders, which are of course the structures of practical engineering interest.

Part I of this paper deals with both plane and coaxial lines having laminated conductors and having the space between the conductors filled with a suitable main dielectric, which may so far as the theory is concerned also be a nonconducting magnetic material. Structures of this type are called "Clogston 1" transmission lines. Although in principle the total space may be divided between the main dielectric and the laminated stacks in any desired ratio, we suppose in Part I that the width of the main dielectric is several times the total thickness of the laminations. When this is true, the principal mode fields in the main dielectric are almost identical to the fields of the transverse electromagnetic (TEM) mode between perfectly conducting planes or cylinders. The phase velocity is controlled by the properties of the main dielectric, while the attenuation constant is determined by the surface impedances of the laminated boundaries (and the dissipation, if any, in the main dielectric). The calculation of the surface impedance of a laminated plane or cylindrical stack is reduced, using the generalized impedance concept developed by Schelkunoff, to the calculation of the input impedance of a chain of transducers with known impedance elements, the chain also being terminated in a known impedance. We are thus able to employ the language and the results of one-dimensional transmission theory to solve our three-dimensional field problem.

In the remaining sections of Part I we introduce various simplifying approximations and special assumptions into the general equations in order to obtain simple and explicit results. We first calculate the propagation constant and the field components of the principal mode under the assumption that the individual conducting laminae are extremely thin compared to the skin depth at the operating frequency, and show that the attenuation constant is substantially independent of frequency so long as this assumption is valid. We then give formulas for the reduction of the effective skin depth in the stacks and the consequent increase of attenuation with frequency when the laminae are of finite thickness. Next we investigate the effect of varying the phase velocity of the line away from the optimum value given by Clogston; and in the last section

we discuss losses due to imperfect dielectrics and lossy magnetic materials.

Part II will be largely devoted to transmission lines of the so-called "Clogston 2" type, in which the entire propagation space is filled with the laminated medium, though to a lesser extent we shall also consider transmission lines having an arbitrary fraction of their total volume filled with laminations and the rest with dielectric. We shall first derive expressions for the propagation constant and the fields of the lowest Clogston 2 mode assuming infinitesimally thin laminae, so that the attenuation constant is essentially independent of frequency, and then go on to investigate the transition of the lowest Clogston 1 mode into the lowest Clogston 2 mode as the space occupied by the main dielectric is gradually filled with laminations. We shall also discuss the higher modes which can exist in Clogston 1 and Clogston 2 lines with infinitesimally thin laminae. Next the effect of finite lamina thickness on the variation of attenuation with frequency in a Clogston 2 will be investigated, and then the important question of the influence of nonuniformity of the laminated medium on the transmission properties of the line. We shall conclude with a short section on dielectric and magnetic losses.

Insofar as possible, plane and coaxial lines will be treated together throughout the paper. Since however Bessel functions are not so easy to manipulate as hyperbolic functions, there will be a few cases where explicit formulas are not yet available for the cylindrical geometry. In these cases the formulas derived for the parallel-plane geometry usually provide reasonably good approximations, or if greater accuracy is desired specific examples may be worked out numerically from the fundamental equations in cylindrical coordinates.

The purpose of the present paper is to set up a general mathematical framework for the analysis of laminated transmission lines, and to treat the major theoretical questions which arise in connection with these lines. In view of the length of the mathematical analysis, we have not devoted much space to numerical examples, although a large number of specific formulas are given which may be used to calculate the theoretical performance of almost any Clogston-type line that happens to be of interest. A considerable part of our work is directed toward evaluating the effects of deviations from the ideal Clogston structure. Both theoretical and experimental results suggest that the limitations on the ultimate applications of the Clogston cable are likely to be imposed by practical problems of manufacture. These limitations, however, depend upon engineering questions which we shall not consider here.

II. WAVE PROPAGATION BETWEEN PLANE AND CYLINDRICAL IMPEDANCE SHEETS

We shall consider waves in a homogeneous, isotropic medium of dielectric constant ϵ , permeability μ , and conductivity g (rationalized MKS units). When convenient we shall also describe the medium in terms of the secondary electromagnetic constants σ and η , defined by

$$\sigma = \sqrt{i\omega\mu(g + i\omega\epsilon)}, \quad \eta = \sqrt{i\omega\mu/(g + i\omega\epsilon)}. \quad (1)$$

The quantity σ is called the intrinsic propagation constant and η the intrinsic impedance of the medium.

We begin by considering structures bounded by infinite planes parallel to the x - z coordinate plane, and we confine our attention to transverse magnetic waves propagating in the z -direction. We assume that the only non-vanishing component of magnetic field is H_x , and that all the fields are independent of x . Then the non-zero field components, written to indicate their dependence on the spatial coordinates, are $H_x(y, z)$, $E_y(y, z)$ and $E_z(y, z)$, the time dependence $e^{i\omega t}$ being understood throughout. The field components are shown in Fig. 1.

The field vectors are connected by Maxwell's two curl equations, which reduce in the present case to

$$\begin{aligned} \partial H_x / \partial z &= (g + i\omega\epsilon)E_y, \\ \partial H_x / \partial y &= -(g + i\omega\epsilon)E_z, \end{aligned} \quad (2)$$

and

$$\partial E_y / \partial z - \partial E_z / \partial y = i\omega\mu H_x. \quad (3)$$

If we eliminate E_y and E_z we get

$$\partial^2 H_x / \partial y^2 + \partial^2 H_x / \partial z^2 = \sigma^2 H_x, \quad (4)$$

where σ is the intrinsic propagation constant defined above. It is easy to see that (4) is satisfied by a wave function of exponential form, say

$$H_x = e^{-\kappa y - \gamma z}, \quad (5)$$

provided that the constants κ and γ are such that

$$\kappa^2 + \gamma^2 = \sigma^2. \quad (6)$$

We may regard κ and γ as the (possibly complex) propagation constants in the y - and z -directions respectively. Either may be chosen at will and the other is then determined by the condition (6). The electric field com-

ponents corresponding to any particular H_x are easily obtained from equations (2).

A concept important in what follows is that of wave impedances³ at a point. For a wave whose field components are H_x , E_y , E_z , the wave impedances looking in the positive and negative y - and z -directions at a typical point are defined to be, respectively,

$$\begin{aligned} Z_y^+ &= E_z/H_x, & Z_z^+ &= -E_y/H_x, \\ Z_y^- &= -E_z/H_x, & Z_z^- &= E_y/H_x. \end{aligned} \quad (7)$$

For waves of the type that we consider, Z_y^+ and Z_y^- are functions of y only, so that if two media having different electrical properties are separated by the plane $y = y_0$, the continuity of the tangential compo-

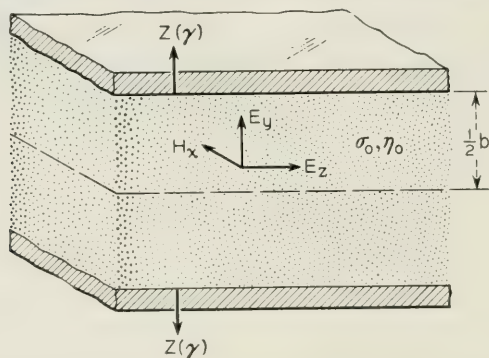


Fig. 1—Transmission line bounded by parallel impedance sheets.

nents of \mathbf{E} and \mathbf{H} across the boundary can be assured by merely requiring the continuity of Z_y^+ (say) at $y = y_0$. This is equivalent to the requirement that the sum of the impedances Z_y^+ and Z_y^- looking into the media on opposite sides of the boundary be zero. A similar condition holds for the impedances Z_z^+ and Z_z^- at a boundary $z = z_0$.

As an example of the use of the wave impedance concept, we shall consider the propagation of a transverse magnetic wave between parallel impedance sheets⁴ which are separated by a distance b . For the moment nothing is specified about the structure of the sheets except that the normal surface impedance looking into each is $Z(\gamma)$, for a wave whose propagation constant in the z -direction is γ . The fact that in general Z will depend upon γ should be noted, since in some cases this dependence

³ S. A. Schelkunoff, *Electromagnetic Waves*, D. van Nostrand Co., Inc., New York, 1943, pp. 249–251. Since in our problem three field components vanish identically, we need only two of the six impedances which are defined in the general case.

⁴ Reference 3, pp. 484–489.

is quite important. The sheets are located at $y = \pm \frac{1}{2}b$, as shown in Fig. 1, and the space between them is filled with a medium whose electrical constants are ϵ_0 , μ_0 , g_0 (or σ_0 , η_0 , if we wish to use the derived constants).

From the symmetry of the boundary conditions it is evident that for any particular mode H_x must be either an even function or an odd function of y about the plane $y = 0$. Taking the even case first, we have

$$\begin{aligned} H_x &= \text{ch } \kappa_0 y \, e^{-\gamma z}, \\ E_y &= -\frac{\gamma}{g_0 + i\omega\epsilon_0} \text{ch } \kappa_0 y \, e^{-\gamma z}, \\ E_z &= -\frac{\kappa_0}{g_0 + i\omega\epsilon_0} \text{sh } \kappa_0 y \, e^{-\gamma z}, \end{aligned} \quad (8)$$

where

$$\kappa_0^2 + \gamma^2 = \sigma_0^2. \quad (9)$$

If we replace $g_0 + i\omega\epsilon_0$ by σ_0/η_0 and κ_0 by $(\sigma_0^2 - \gamma^2)^{\frac{1}{2}}$, the boundary condition at $y = \frac{1}{2}b$, namely

$$Z_y^+ = Z(\gamma), \quad (10)$$

becomes

$$\frac{1}{2}(\sigma_0^2 - \gamma^2)^{\frac{1}{2}}b \tanh \frac{1}{2}(\sigma_0^2 - \gamma^2)^{\frac{1}{2}}b = -\frac{\sigma_0 b}{2\eta_0} Z(\gamma). \quad (11)$$

Similarly, the odd case gives

$$\begin{aligned} H_x &= \text{sh } \kappa_0 y \, e^{-\gamma z}, \\ E_y &= -\frac{\gamma}{g_0 + i\omega\epsilon_0} \text{sh } \kappa_0 y \, e^{-\gamma z}, \\ E_z &= -\frac{\kappa_0}{g_0 + i\omega\epsilon_0} \text{ch } \kappa_0 y \, e^{-\gamma z}; \end{aligned} \quad (12)$$

and the boundary condition becomes

$$\frac{1}{2}(\sigma_0^2 - \gamma^2)^{\frac{1}{2}}b \coth \frac{1}{2}(\sigma_0^2 - \gamma^2)^{\frac{1}{2}}b = -\frac{\sigma_0 b}{2\eta_0} Z(\gamma). \quad (13)$$

The transcendental equations (11) and (13) are satisfied by the propagation constants of the various even and odd modes; presumably each has an infinite number of roots, which we could find, at least in principle, if we knew the explicit form of the function $Z(\gamma)$. We shall confine ourselves here to deriving an approximate expression for the propagation

constant of the principal mode (lowest even mode) when the walls are very good conductors.

If the walls were perfectly conducting we should have $Z(\gamma) = 0$, and the lowest root γ_0 of (11) would be given by

$$(\sigma_0^2 - \gamma_0^2)^{\frac{1}{2}}b = 0, \quad \text{or} \quad \gamma_0 = \sigma_0. \quad (14)$$

The principal mode between perfectly conducting sheets is just an undisturbed slice of the plane TEM wave which could propagate in an unbounded medium. If $Z(\gamma_0)$ is not rigorously zero, but still so small that

$$\left| \frac{\sigma_0 b Z(\gamma_0)}{2\eta_0} \right| \ll 1, \quad (15)$$

and if $Z(\gamma)$ does not vary rapidly with γ in the neighborhood of γ_0 , then the lowest root of (11) is given approximately by

$$\gamma^2 = \sigma_0^2 + 2\sigma_0 Z(\gamma_0)/\eta_0 b. \quad (16)$$

If $Z(\gamma_0)$ is so small that we have the further inequality

$$\frac{1}{2} \left| \frac{Z(\gamma_0)}{\sigma_0 b \eta_0} \right|^2 \ll 1, \quad (17)$$

then (16) yields the approximation

$$\gamma = \sigma_0 + Z(\gamma_0)/\eta_0 b, \quad (18)$$

where the second term is the first-order change in γ due to the finite impedance of the walls. If we formally set $g_0 = 0$ (this does not actually restrict us to perfect dielectrics since we could still assume ϵ_0 or μ_0 to be complex), we have

$$\sigma = i\omega\sqrt{\mu_0\epsilon_0}, \quad \eta = \sqrt{\mu_0/\epsilon_0}. \quad (19)$$

If the medium between the sheets is lossless, the attenuation and phase constants of the principal mode become

$$\alpha = \text{Re } \gamma = \text{Re } Z(\gamma_0)/\eta_0 b, \quad (20)$$

$$\beta = \text{Im } \gamma = \omega\sqrt{\mu_0\epsilon_0} + \text{Im } Z(\gamma_0)/\eta_0 b. \quad (21)$$

Although the fields of the principal mode between perfectly conducting walls are entirely transverse to the direction of propagation, if the walls are not perfectly conducting there will also be a small longitudinal component E_z of electric field associated with this mode. The leading terms in the expressions for the field components, as obtained from equations (8), (9), and (16), are

$$\begin{aligned}
H_x &\approx H_0 e^{-\gamma z}, \\
E_y &\approx -\eta_0 H_0 e^{-\gamma z}, \\
E_z &\approx \frac{2Z_0(\gamma_0)H_0 y}{b} e^{-\gamma z},
\end{aligned}
\tag{22}$$

where H_0 is an arbitrary amplitude factor.

As an example of the use of (20) and (21), suppose that the impedance sheets in Fig. 1 are electrically thick metal walls of permeability μ_1 and (high) conductivity g_1 . Then to a very good approximation at all engineering frequencies and for all ordinary dielectrics between the walls, the surface impedance is

$$Z(\gamma_0) = (1 + i)/g_1 \delta_1, \tag{23}$$

where

$$\delta_1 = \sqrt{2/\omega \mu_1 g_1} \tag{24}$$

is the skin depth in the metal. We thus obtain from (20) and (21) the familiar formulas

$$\alpha = 1/\eta_0 b g_1 \delta_1, \tag{25}$$

$$\beta = \omega \sqrt{\mu_0 \epsilon_0} + 1/\eta_0 b g_1 \delta_1. \tag{26}$$

It should be noted that in practical cases the inequality (17) on which we based the approximations (20) and (21) does not hold down to the mathematical limit of zero frequency. In the present paper, however, when we speak of "low frequencies" we shall mean frequencies still high enough so that the approximations (20) and (21) for α and β are valid. Generally this will be equivalent to the assumption that the attenuation per radian is small. In our applications this assumption will usually be justified down to frequencies of the order of a few $\text{kc} \cdot \text{sec}^{-1}$.

Now let us consider transmission lines bounded by coaxial circular cylinders and confine our attention to circular transverse magnetic waves propagating in the z -direction. For these waves the fields are independent of the angle ϕ , and the only non-vanishing field components are $H_\phi(\rho, z)$, $E_\rho(\rho, z)$, and $E_z(\rho, z)$. The field components are shown in Fig. 2.

For circular transverse magnetic fields Maxwell's curl equations in a homogeneous, isotropic medium reduce to

$$\begin{aligned}
\partial H_\phi / \partial z &= -(g + i\omega\epsilon) E_\rho, \\
\partial(\rho H_\phi) / \partial \rho &= (g + i\omega\epsilon) \rho E_z,
\end{aligned}
\tag{27}$$

and

$$\partial E_z / \partial \rho - \partial E_\rho / \partial z = i\omega\mu H_\phi, \quad (28)$$

from which we can eliminate E_ρ and E_z to obtain

$$\frac{\partial^2 H_\phi}{\partial \rho^2} + \frac{1}{\rho} \frac{\partial H_\phi}{\partial \rho} - \frac{H_\phi}{\rho^2} + \frac{\partial^2 H_\phi}{\partial z^2} = \sigma^2 H_\phi. \quad (29)$$

If we assume a wave traveling in the positive z -direction with propagation constant γ and write

$$H_\phi(\rho, z) = R(\rho)e^{-\gamma z}, \quad (30)$$

we find that (29) becomes

$$\frac{1}{\rho} \frac{d}{d\rho} \left(\rho \frac{dR}{d\rho} \right) - \left(\kappa^2 + \frac{1}{\rho^2} \right) R = 0, \quad (31)$$

where κ is given by (6) as before. But (31) is just the equation satisfied by modified Bessel functions of order one and argument $\kappa\rho$, so

$$R(\rho) = AI_1(\kappa\rho) + BK_1(\kappa\rho), \quad (32)$$

where A and B are arbitrary constants. The other field components can be obtained from H_ϕ using (27); the results are

$$H_\phi = [AI_1(\kappa\rho) + BK_1(\kappa\rho)]e^{-\gamma z},$$

$$E_\rho = \frac{\gamma}{g + i\omega\epsilon} [AI_1(\kappa\rho) + BK_1(\kappa\rho)]e^{-\gamma z}, \quad (33)$$

$$E_z = \frac{\kappa}{g + i\omega\epsilon} [AI_0(\kappa\rho) - BK_0(\kappa\rho)]e^{-\gamma z}.$$

For cylindrical fields of the type that we are considering, the wave impedances looking in the positive and negative ρ - and z -directions at a typical point are defined to be, respectively,

$$\begin{aligned} Z_\rho^+ &= -E_z / H_\phi, & Z_z^+ &= E_\rho / H_\phi, \\ Z_\rho^- &= E_z / H_\phi, & Z_z^- &= -E_\rho / H_\phi. \end{aligned} \quad (34)$$

We shall now discuss the propagation of circular transverse magnetic waves in a homogeneous region of space whose electrical constants are ϵ_0 , μ_0 , g_0 (or σ_0 , η_0), and which is bounded by coaxial cylinders of radii ρ_1 and ρ_2 , where $\rho_2 > \rho_1$, as shown in Fig. 2. We suppose that the radial impedances looking from the main dielectric into the inner and outer cylinders are, respectively,

$$Z_\rho^-|_{\rho=\rho_1} = Z_1(\gamma), \quad Z_\rho^+|_{\rho=\rho_2} = Z_2(\gamma). \quad (35)$$

Then from (33) and (34) the boundary conditions are

$$\begin{aligned} \eta_{0\rho} \frac{AI_0(\kappa_0\rho_1) - BK_0(\kappa_0\rho_1)}{AI_1(\kappa_0\rho_1) + BK_1(\kappa_0\rho_1)} &= Z_1(\gamma), \\ \eta_{0\rho} \frac{AI_0(\kappa_0\rho_2) - BK_0(\kappa_0\rho_2)}{AI_1(\kappa_0\rho_2) + BK_1(\kappa_0\rho_2)} &= -Z_2(\gamma), \end{aligned} \quad (36)$$

where

$$\kappa_0 = (\sigma_0^2 - \gamma^2)^{\frac{1}{2}}, \quad \eta_{0\rho} = \frac{\kappa_0}{g_0 + i\omega\epsilon_0} = \eta_0(1 - \gamma^2/\sigma_0^2)^{\frac{1}{2}}. \quad (37)$$

If equations (36) are to be satisfied by values of A and B which are not both zero, it is easily shown that a necessary and sufficient condition is

$$\frac{\eta_{0\rho}K_0(\kappa_0\rho_1) + Z_1(\gamma)K_1(\kappa_0\rho_1)}{\eta_{0\rho}I_0(\kappa_0\rho_1) - Z_1(\gamma)I_1(\kappa_0\rho_1)} = \frac{\eta_{0\rho}K_0(\kappa_0\rho_2) - Z_2(\gamma)K_1(\kappa_0\rho_2)}{\eta_{0\rho}I_0(\kappa_0\rho_2) + Z_2(\gamma)I_1(\kappa_0\rho_2)}, \quad (38)$$

and (38) is a transcendental equation for the determination of the propagation constants of all the circular magnetic modes in the coaxial line.

As in the discussion of the parallel-plane line, we shall confine our attention to the principal mode and shall assume forthwith that the wall losses are small.⁵ Since for the principal mode we expect that γ will be nearly equal to σ_0 , we may write γ_0 for σ_0 and evaluate Z_1 and Z_2 at γ_0 ; and we may replace the modified Bessel functions in (38) by their ap-

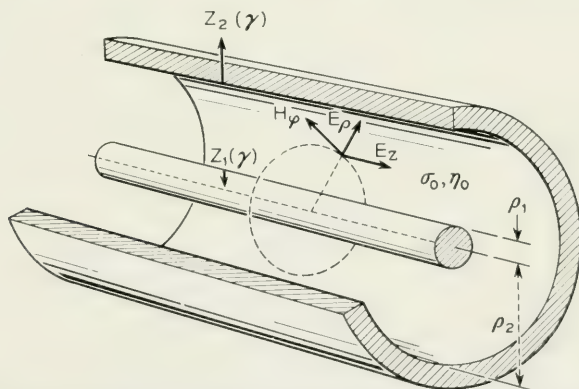


Fig. 2—Transmission line bounded by coaxial impedance cylinders.

⁵ J. A. Stratton, *Electromagnetic Theory*, McGraw-Hill, New York, 1941, pp. 551-554, gives a similar treatment of the principal mode in an ordinary coaxial cable with solid metal walls.

proximate values for small argument. From the series given in Dwight⁶ 813.1, 813.2, 815.1, and 815.2, we have

$$\begin{aligned} I_0(x) &\approx 1, \\ I_1(x) &\approx \frac{1}{2}x, \\ K_0(x) &\approx -(0.5772 + \log \frac{1}{2}x) = -\log 0.8905x, \\ K_1(x) &\approx \frac{1}{x} + \frac{1}{2}x \log 0.8905x, \end{aligned} \quad (39)$$

for $|x| \ll 1$, where \log represents the natural logarithm. If we put these approximations into (38) and if we suppose that the wall impedances are so small that

$$|\sigma_0 \rho_1 Z_1(\gamma_0)/2\eta_0| \ll 1, \quad |\sigma_0 \rho_2 Z_2(\gamma_0)/2\eta_0| \ll 1, \quad (40)$$

we obtain, after a little algebra,

$$\kappa_0^2 = \sigma_0^2 - \gamma^2 = -\frac{\sigma_0[Z_1(\gamma_0)/\rho_1 + Z_2(\gamma_0)/\rho_2]}{\eta_0 \log(\rho_2/\rho_1)}. \quad (41)$$

Now further assuming that

$$\frac{1}{8} \left| \frac{Z_1(\gamma_0)/\rho_1 + Z_2(\gamma_0)/\rho_2}{\sigma_0 \eta_0 \log(\rho_2/\rho_1)} \right|^2 \ll 1, \quad (42)$$

we get by the binomial theorem

$$\gamma = \sigma_0 + \frac{Z_1(\gamma_0)/\rho_1 + Z_2(\gamma_0)/\rho_2}{2\eta_0 \log(\rho_2/\rho_1)}. \quad (43)$$

If we formally set $g_0 = 0$, we find that the attenuation and phase constants of the principal mode in a coaxial line with low-loss walls and no dissipation in the main dielectric are

$$\alpha = \operatorname{Re} \gamma = \operatorname{Re} \frac{Z_1(\gamma_0)/\rho_1 + Z_2(\gamma_0)/\rho_2}{2\eta_0 \log(\rho_2/\rho_1)}, \quad (44)$$

$$\beta = \operatorname{Im} \gamma = \omega \sqrt{\mu_0 \epsilon_0} + \operatorname{Im} \frac{Z_1(\gamma_0)/\rho_1 + Z_2(\gamma_0)/\rho_2}{2\eta_0 \log(\rho_2/\rho_1)}. \quad (45)$$

As before, these approximations for α and β will ultimately break down as the frequency approaches zero, but they will certainly be valid over the frequency range in which we are interested in the present paper.

⁶ H. B. Dwight, *Tables of Integrals and Other Mathematical Data*, Revised Edition, Macmillan, New York, 1947. We shall refer to Dwight for a number of standard series expansions.

The magnetic field lines of the principal mode will of course be circles and the electric field will be largely radial, but with a small longitudinal component unless the wall impedances are rigorously zero. The general expressions (33) for the fields may be reduced to simple approximate formulas if we use the fact that κ_0^2 is given by (41) and $\kappa_0\rho$ is small compared to unity. The ratio A/B may be obtained from either of equations (36). Introducing the approximations (39) for the Bessel functions and carrying out a little algebra, we get the following approximate expressions for the fields:

$$\begin{aligned} H_\phi &\approx \frac{I}{2\pi\rho} e^{-\gamma z}, \\ E_\rho &\approx \frac{\eta_0 I}{2\pi\rho} e^{-\gamma z}, \\ E_z &\approx \frac{I}{2\pi \log(\rho_2/\rho_1)} \left[\frac{Z_1(\gamma_0)}{\rho_1} \log \frac{\rho_2}{\rho} + \frac{Z_2(\gamma_0)}{\rho_2} \log \frac{\rho_1}{\rho} \right] e^{-\gamma z}, \end{aligned} \quad (46)$$

where the amplitude factor I is equal to the total current flowing in the inner cylinder. Incidentally we note that the above results might have been derived from more elementary arguments if we had started with the fields in a coaxial line with perfectly conducting walls and treated the effect of finite wall impedance as a small perturbation.

If we consider an ordinary coaxial cable with solid metal walls at a frequency high enough so that there is a well-developed skin effect on both conductors, then to a good approximation

$$Z_1(\gamma_0) = Z_2(\gamma_0) = (1 + i)/g_1\delta_1, \quad (47)$$

where g_1 and δ_1 are the conductivity and the skin thickness of the metal; and the attenuation and phase constants are given by the well-known expressions

$$\alpha = \frac{1/\rho_1 + 1/\rho_2}{2\eta_0 g_1 \delta_1 \log(\rho_2/\rho_1)}, \quad (48)$$

$$\beta = \omega\sqrt{\mu_0\epsilon_0} + \frac{1/\rho_1 + 1/\rho_2}{2\eta_0 g_1 \delta_1 \log(\rho_2/\rho_1)}. \quad (49)$$

If necessary we may take account of dissipation in the main dielectric of either a plane or a coaxial transmission line by assigning complex values⁷ to ϵ_0 and μ_0 , say

⁷ See, for example, C. G. Montgomery, *Principles of Microwave Circuits*, M. I. T. Rad. Lab. Series, **8**, McGraw-Hill, New York, 1948, pp. 365-369 and 382-385.

$$\begin{aligned}\epsilon_0 &= \epsilon'_0 - i\epsilon''_0 = \epsilon'_0(1 - i \tan \phi_0), \\ \mu_0 &= \mu'_0 - i\mu''_0 = \mu'_0(1 - i \tan \zeta_0),\end{aligned}\quad (50)$$

where $\tan \phi_0$ is the dielectric loss tangent and $\tan \zeta_0$ is the magnetic loss tangent (if any). Inserting (50) into (18) or (43), we find for the attenuation due to dielectric and magnetic losses,

$$\begin{aligned}\alpha_d &= \text{Re } \sigma = \text{Re } i\omega \sqrt{\mu'_0 \epsilon'_0 (1 - i \tan \phi_0)(1 - i \tan \zeta_0)} \\ &= \frac{1}{2} \omega \sqrt{\mu'_0 \epsilon'_0} (\tan \phi_0 + \tan \zeta_0),\end{aligned}\quad (51)$$

provided that $\tan \phi_0$ and $\tan \zeta_0$ are both small compared to unity, as they will always be in practice. We shall neglect second-order effects and so regard the dielectric losses, the magnetic losses, and the wall losses as additive.

III. SURFACE IMPEDANCE OF A LAMINATED BOUNDARY

The main problem in the theory of Clogston 1 transmission lines is the computation of the surface impedance of a laminated plane or cylindrical boundary having alternate thin layers of conductor and dielectric. Portions of such laminated structures are shown schematically in Figs. 3 and 4. We shall begin with an analysis, similar to Clogston's,⁸ of the plane stack. This will lead to a convenient point of view for the treatment of the mathematically more complicated coaxial stack.

Let us consider a wave with field components H_x , E_y , E_z , propagating

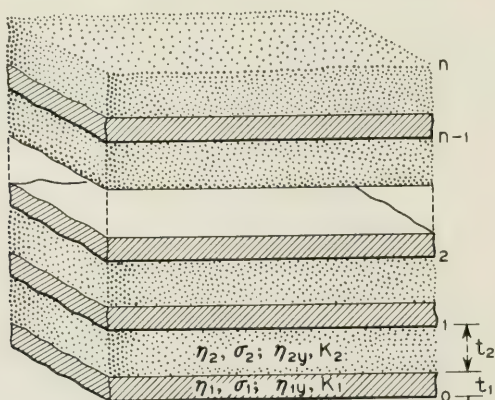


Fig. 3—Portion of laminated plane stack.

⁸ Reference 1, Section III.

in a layer of homogeneous, isotropic material whose electrical constants are ϵ , μ , g (or σ , η), and which is bounded by planes perpendicular to the y -axis. Henceforth we shall always assume that the z -dependence of every field component is given by the factor $e^{-\gamma z}$, where the complex quantity γ , whose value may or may not be known a priori, is the propagation constant of the wave in the z -direction. Then the first of Maxwell's equations (2) yields

$$E_y = -[\gamma/(g + i\omega\epsilon)]H_x, \quad (52)$$

and on eliminating E_y from the other Maxwell equations, we get

$$\begin{aligned} \partial H_x / \partial y &= -(g + i\omega\epsilon)E_z, \\ \partial E_z / \partial y &= -[\kappa^2/(g + i\omega\epsilon)]H_x, \end{aligned} \quad (53)$$

where κ^2 is defined by equation (6).

Now if we formally identify H_x with "current" and E_z with "voltage", equations (53) are just the equations of a uniform one-dimensional transmission line extending in the y -direction, with series impedance $\kappa^2/(g + i\omega\epsilon)$ per unit length and shunt admittance $(g + i\omega\epsilon)$ per unit length; in other words a transmission line whose propagation constant is κ and whose characteristic impedance is η_y , where

$$\kappa = \sigma(1 - \gamma^2/\sigma^2)^{\frac{1}{2}}, \quad \eta_y = \kappa/(g + i\omega\epsilon) = \eta(1 - \gamma^2/\sigma^2)^{\frac{1}{2}}. \quad (54)$$

Hence we can apply the whole theory of one-dimensional transmission lines with the assurance that in so doing we shall not violate the field equations. For example, if $E(0)$, $H(0)$ and $E(t)$, $H(t)$ represent the tangential field components E_z , H_x at two planes separated by a dis-

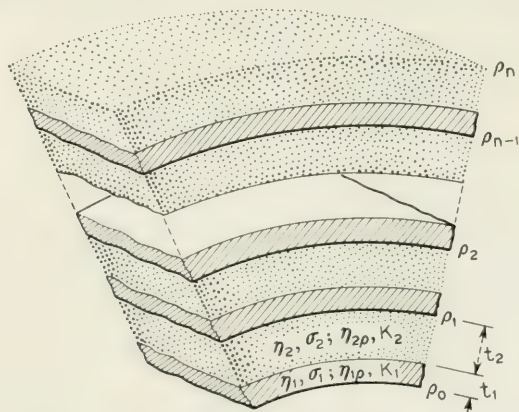


Fig. 4—Portion of laminated coaxial stack.

tance t , these fields are related by the general circuit parameter matrix of a uniform line, namely

$$\begin{pmatrix} E(0) \\ H(0) \end{pmatrix} = \begin{pmatrix} \text{ch } \kappa t & \eta_y \text{ sh } \kappa t \\ \frac{\text{sh } \kappa t}{\eta_y} & \text{ch } \kappa t \end{pmatrix} \begin{pmatrix} E(t) \\ H(t) \end{pmatrix}. \quad (55)$$

We are now in a position to determine the surface impedance normal to a laminated plane structure composed of layers of which every other one has thickness t_1 and electrical constants σ_1 , η_1 , while the intervening layers each have thickness t_2 and electrical constants σ_2 , η_2 . Fig. 3 shows the cross section of such a stack in which the total number of double layers is n ($2n$ single layers), while Fig. 4 represents the corresponding coaxial stack. Ultimately we shall assume the layers of thickness t_1 to be good conductors and those of thickness t_2 to be good insulators, but these assumptions need not be brought in immediately.

If the fields in the plane stack all vary with z according to $e^{-\gamma z}$, then when we look in the direction of increasing y each double layer may be regarded as a four-terminal network formed by two sections of uniform transmission line of lengths t_1 and t_2 , the propagation constants and characteristic impedances of the two sections being given respectively by

$$\begin{aligned} \kappa_1 &= \sigma_1(1 - \gamma^2/\sigma_1^2)^{\frac{1}{2}}, & \eta_{1y} &= \eta_1(1 - \gamma^2/\sigma_1^2)^{\frac{1}{2}}, \\ \kappa_2 &= \sigma_2(1 - \gamma^2/\sigma_2^2)^{\frac{1}{2}}, & \eta_{2y} &= \eta_2(1 - \gamma^2/\sigma_2^2)^{\frac{1}{2}}. \end{aligned} \quad (56)$$

The matrix of the double layer is the product of the matrices of the two single layers in the proper order. Thus if the tangential field components are E_0 , H_0 at the lower surface of the first layer and E_1 , H_1 at the upper surface of the second layer, we have

$$\begin{pmatrix} E_0 \\ H_0 \end{pmatrix} = \begin{pmatrix} \mathfrak{A} & \mathfrak{B} \\ \mathfrak{C} & \mathfrak{D} \end{pmatrix} \begin{pmatrix} E_1 \\ H_1 \end{pmatrix}, \quad (57)$$

where

$$\begin{aligned} \mathfrak{A} &= \text{ch } \kappa_1 t_1 \text{ ch } \kappa_2 t_2 + \frac{\eta_{1y}}{\eta_{2y}} \text{ sh } \kappa_1 t_1 \text{ sh } \kappa_2 t_2, \\ \mathfrak{B} &= \eta_{2y} \text{ ch } \kappa_1 t_1 \text{ sh } \kappa_2 t_2 + \eta_{1y} \text{ sh } \kappa_1 t_1 \text{ ch } \kappa_2 t_2, \\ \mathfrak{C} &= \frac{1}{\eta_{1y}} \text{ sh } \kappa_1 t_1 \text{ ch } \kappa_2 t_2 + \frac{1}{\eta_{2y}} \text{ ch } \kappa_1 t_1 \text{ sh } \kappa_2 t_2, \\ \mathfrak{D} &= \frac{\eta_{2y}}{\eta_{1y}} \text{ sh } \kappa_1 t_1 \text{ sh } \kappa_2 t_2 + \text{ch } \kappa_1 t_1 \text{ ch } \kappa_2 t_2. \end{aligned} \quad (58)$$

The stack of double layers may be regarded as a chain of iterated four-poles; such chains have an extensive literature.⁹ The relation between the tangential fields E_n , H_n at the upper surface of the n th double layer and E_0 , H_0 at the lower surface of the first double layer is

$$\begin{pmatrix} E_0 \\ H_0 \end{pmatrix} = \mathbf{M}^n \begin{pmatrix} E_n \\ H_n \end{pmatrix}, \quad (59)$$

where \mathbf{M} is the \mathcal{ACD} -matrix appearing in equation (57). However there is a simple expression¹⁰ for the n th power of a square matrix of order two, namely

$$\mathbf{M}^n = M^{\frac{1}{2}(n-1)} \frac{\text{sh } n\Gamma}{\text{sh } \Gamma} \mathbf{M} - M^{\frac{1}{2}n} \frac{\text{sh } (n-1)\Gamma}{\text{sh } \Gamma} \mathbf{I}, \quad (60)$$

where \mathbf{I} is the unit matrix of order two, Γ is the propagation constant per section of the chain of four-poles, defined by

$$\text{ch } \Gamma = (\mathcal{A} + \mathcal{D})/2M^{\frac{1}{2}}, \quad (61)$$

and M is the determinant of the matrix \mathbf{M} , that is,

$$M = \mathcal{A}\mathcal{D} - \mathcal{B}\mathcal{C}. \quad (62)$$

The determinant of the matrix whose elements are given by (58) is unity, as may easily be verified; but this may not be the case for all the matrices which occur in our study of cylindrical structures. M will therefore be carried explicitly in the following equations.

We now introduce the iterative impedances K_1 and K_2 , defined by

$$\begin{aligned} K_1 &= \frac{(\mathcal{A} - \mathcal{D}) + \sqrt{(\mathcal{A} + \mathcal{D})^2 - 4M}}{2\mathcal{C}}, \\ K_2 &= \frac{-(\mathcal{A} - \mathcal{D}) + \sqrt{(\mathcal{A} + \mathcal{D})^2 - 4M}}{2\mathcal{C}}. \end{aligned} \quad (63)$$

K_1 is the impedance seen when we look into a semi-infinite stack of double layers if the first layer is of type 1, while K_2 is the impedance seen if the first layer is of type 2. In calculations relating to Clogston 1 lines with dissipative walls, the real parts of K_1 and K_2 will both be positive. By a straightforward procedure we may express the matrix elements \mathcal{A} , \mathcal{B} , \mathcal{C} , \mathcal{D} in terms of K_1 , K_2 , Γ , and M , and then transform equation

⁹ See, for example, E. A. Guillemin, *Communication Networks*, **2**, Wiley, New York, 1935, pp. 161-166.

¹⁰ F. Abelès, *Comptes Rendus*, **226**, 1872 (1948). This result was called to the author's attention by Mr. J. G. Kreer.

(60) into

$$\mathbf{M}^n = \frac{2M^{\frac{1}{2}n}}{(K_1 + K_2)} \begin{pmatrix} \frac{1}{2}(K_1 e^{n\Gamma} + K_2 e^{-n\Gamma}) & K_1 K_2 \operatorname{sh} n\Gamma \\ \operatorname{sh} n\Gamma & \frac{1}{2}(K_1 e^{-n\Gamma} + K_2 e^{n\Gamma}) \end{pmatrix}. \quad (64)$$

Finally we obtain from (59) and (64) an expression for the impedance Z_0 looking into a plane stack of n double layers when the n th layer is backed by a surface whose impedance is Z_n , namely

$$Z_0 = \frac{E_0}{H_0} = \frac{\frac{1}{2}Z_n(K_1 e^{n\Gamma} + K_2 e^{-n\Gamma}) + K_1 K_2 \operatorname{sh} n\Gamma}{Z_n \operatorname{sh} n\Gamma + \frac{1}{2}(K_1 e^{-n\Gamma} + K_2 e^{n\Gamma})}. \quad (65)$$

For the cylindrical geometry, matters are a good deal more complicated. If we consider waves having field components H_ϕ , E_ρ , E_z in a homogeneous, isotropic shell bounded by coaxial cylindrical surfaces, and assume a propagation factor $e^{-\gamma z}$, Maxwell's equations (27) and (28) may be written

$$E_\rho = [\gamma/(g + i\omega\epsilon)]H_\phi, \quad (66)$$

and

$$\begin{aligned} \partial(-\rho H_\phi)/\partial\rho &= -(g + i\omega\epsilon)\rho E_z, \\ \partial E_z/\partial\rho &= -[\kappa^2/(g + i\omega\epsilon)\rho](-\rho H_\phi). \end{aligned} \quad (67)$$

If desired, we might identify E_z with "voltage" and $-\rho H_\phi$ with "current" and regard equations (67) as describing a nonuniform radial transmission line, having series impedance $\kappa^2/(g + i\omega\epsilon)\rho$ per unit length and shunt admittance $(g + i\omega\epsilon)\rho$ per unit length. Since, however, in equations (34) we have already defined the radial wave impedance to be a field ratio without the extra factor of ρ , we shall carry out the analysis of the present paper directly in terms of the field components E_z and $-H_\phi$.

From the general expressions (33) for the fields in cylindrical coordinates, we can show that the matrix relation between the tangential field components E_z , $-H_\phi$ at two radii ρ_1 and ρ_2 is given by

$$\begin{pmatrix} E(\rho_1) \\ -H(\rho_1) \end{pmatrix} = \begin{pmatrix} \kappa\rho_2(K_{01}I_{12} + K_{12}I_{01}) & \eta_\rho\kappa\rho_2(K_{01}I_{02} - K_{02}I_{01}) \\ \frac{\kappa\rho_2}{\eta_\rho}(K_{11}I_{12} - K_{12}I_{11}) & \kappa\rho_2(K_{11}I_{02} + K_{02}I_{11}) \end{pmatrix} \begin{pmatrix} E(\rho_2) \\ -H(\rho_2) \end{pmatrix}, \quad (68)$$

where

$$\kappa = (\sigma^2 - \gamma^2)^{\frac{1}{2}}, \quad \eta_\rho = \eta(1 - \gamma^2/\sigma^2)^{\frac{1}{2}}, \quad (69)$$

and we have used the abbreviations

$$I_{rs} = I_r(\kappa\rho_s), \quad K_{rs} = K_r(\kappa\rho_s). \quad (70)$$

It may be verified that the determinant M of the square matrix appearing in (68) is simply

$$M = \rho_2/\rho_1. \quad (71)$$

In principle equation (68) permits us to determine by matrix multiplication the relation between the tangential fields at the inner and outer surfaces of a coaxial double layer, or of a laminated stack of any number of double layers, such as is shown in Fig. 4. The difficulty is that the elements of the matrix of a single layer are not functions only of the electrical properties of the layer and its thickness, but depend in a more complicated way on the inner and outer radii separately. Whereas in the plane case we had merely to take the n th power of a single matrix, we are now faced with the problem of multiplying together n matrices, each of which differs more or less from all the others. An exact expression for the result is practically out of the question; but we can make some reasonable approximations if we assume that each individual layer is thin compared to its mean radius, so that the matrix elements do not change much from one layer to the next.

If the thickness $t (= \rho_2 - \rho_1)$ of a single layer is small compared to ρ_1 , then the Bessel function combinations appearing in (68) may be expanded in series, as shown in Appendix I, and the circuit parameter matrix takes the following approximate form,

$$\begin{pmatrix} \left[1 + \frac{t}{2\rho_1}\right] \text{ch } \kappa t - \frac{1}{2\kappa\rho_1} \text{sh } \kappa t & \eta_\rho \left[1 + \frac{t}{2\rho_1}\right] \text{sh } \kappa t \\ \frac{1}{\eta_\rho} \left[1 + \frac{t}{2\rho_1}\right] \text{sh } \kappa t & \left[1 + \frac{t}{2\rho_1}\right] \text{ch } \kappa t + \frac{1}{2\kappa\rho_1} \text{sh } \kappa t \end{pmatrix}, \quad (72)$$

where terms of the order of t/ρ_1 represent the first-order curvature corrections. If we use the same value of ρ_1 , say $\bar{\rho}$, for both parts of a double layer, then up to first order the elements of the matrix of the double layer become

$$\begin{aligned}
\alpha &= \left[1 + \frac{t_1 + t_2}{2\bar{\rho}} \right] \left[\text{ch } \kappa_1 t_1 \text{ ch } \kappa_2 t_2 + \frac{\eta_{1\rho}}{\eta_{2\rho}} \text{ sh } \kappa_1 t_1 \text{ sh } \kappa_2 t_2 \right] \\
&\quad - \left[\frac{1}{2\kappa_1 \bar{\rho}} \text{ sh } \kappa_1 t_1 \text{ ch } \kappa_2 t_2 + \frac{1}{2\kappa_2 \bar{\rho}} \text{ ch } \kappa_1 t_1 \text{ sh } \kappa_2 t_2 \right], \\
\beta &= \left[1 + \frac{t_1 + t_2}{2\bar{\rho}} \right] \left[\eta_{2\rho} \text{ ch } \kappa_1 t_1 \text{ sh } \kappa_2 t_2 + \eta_{1\rho} \text{ sh } \kappa_1 t_1 \text{ ch } \kappa_2 t_2 \right] \\
&\quad + \left[\frac{\eta_{1\rho}}{2\kappa_2 \bar{\rho}} - \frac{\eta_{2\rho}}{2\kappa_1 \bar{\rho}} \right] \text{ sh } \kappa_1 t_1 \text{ sh } \kappa_2 t_2, \\
\epsilon &= \left[1 + \frac{t_1 + t_2}{2\bar{\rho}} \right] \left[\frac{1}{\eta_{1\rho}} \text{ sh } \kappa_1 t_1 \text{ ch } \kappa_2 t_2 + \frac{1}{\eta_{2\rho}} \text{ ch } \kappa_1 t_1 \text{ sh } \kappa_2 t_2 \right] \\
&\quad + \left[\frac{1}{2\eta_{2\rho} \kappa_1 \bar{\rho}} - \frac{1}{2\eta_{1\rho} \kappa_2 \bar{\rho}} \right] \text{ sh } \kappa_1 t_1 \text{ sh } \kappa_2 t_2, \\
\mathfrak{D} &= \left[1 + \frac{t_1 + t_2}{2\bar{\rho}} \right] \left[\frac{\eta_{2\rho}}{\eta_{1\rho}} \text{ sh } \kappa_1 t_1 \text{ sh } \kappa_2 t_2 + \text{ch } \kappa_1 t_1 \text{ ch } \kappa_2 t_2 \right] \\
&\quad + \left[\frac{1}{2\kappa_1 \bar{\rho}} \text{ sh } \kappa_1 t_1 \text{ ch } \kappa_2 t_2 + \frac{1}{2\kappa_2 \bar{\rho}} \text{ ch } \kappa_1 t_1 \text{ sh } \kappa_2 t_2 \right].
\end{aligned} \tag{73}$$

As in the analogous equations (58) for a plane double layer, the subscripts 1 and 2 refer to the first and second layers respectively.

If we have a stack of double layers in which all the layers of the same kind have the same thickness and same electrical constants, then the only term in (73) which varies from one double layer to the next is the mean radius $\bar{\rho}$. Depending on the circumstances, we may wish to use a single value of $\bar{\rho}$ for the whole stack, or a few different values, or even, if high-speed computing machinery is available to carry out the matrix multiplications, a different value of $\bar{\rho}$ for each double layer. The matrix of the whole stack then becomes a product of powers of as many different matrices as we have chosen values of $\bar{\rho}$. Obviously this method is better adapted to the numerical analysis of special cases than to the general theoretical treatment of a stack whose ratio of outer radius to inner radius is unspecified.

In principle we are now able to compute the normal surface impedance of any laminated plane or coaxial stack at a given frequency provided that we know the electrical constants and the thickness of each layer, the number of layers, the propagation constant γ in the z -direction, and the normal impedance Z_n of the material behind the last layer. Since the general formulas even for plane stacks are quite complicated, however, we shall introduce at this point some very good approximations which will be valid for all of the following work.

Henceforth we shall take the layers of thickness t_1 to be such good conductors that the ratio $\omega\epsilon_1/g_1$ of displacement current to conduction current is negligible in comparison with unity. For metals like copper this is an excellent approximation at even the highest engineering frequencies. Then on introducing the characteristic skin thickness δ_1 , we have for the conducting layers,

$$\begin{aligned}\sigma_1 &= \sqrt{i\omega\mu_1 g_1} = (1+i)/\delta_1, \\ \eta_1 &= \sqrt{i\omega\mu_1/g_1} = (1+i)/g_1\delta_1,\end{aligned}\tag{74}$$

where

$$\delta_1 = \sqrt{2/\omega\mu_1 g_1}.\tag{75}$$

For pure copper the permeability and conductivity are

$$\begin{aligned}\mu_1 &= 1.257 \times 10^{-6} \text{ henrys}\cdot\text{meter}^{-1}, \\ g_1 &= 5.800 \times 10^7 \text{ mhos}\cdot\text{meter}^{-1},\end{aligned}\tag{76}$$

from which we obtain the numerical values

$$\begin{aligned}\sigma_1 &= 1.513 \times 10^4 (1+i)\sqrt{f_{\text{Mc}}} \text{ meters}^{-1}, \\ \eta_1 &= 2.609 \times 10^{-4} (1+i)\sqrt{f_{\text{Mc}}} \text{ ohms},\end{aligned}\tag{77}$$

and

$$\delta_1 = \frac{6.609 \times 10^{-5}}{\sqrt{f_{\text{Mc}}}} \text{ meters} = \frac{2.602}{\sqrt{f_{\text{Mc}}}} \text{ mils},\tag{78}$$

where f_{Mc} is the frequency in $\text{Mc}\cdot\text{sec}^{-1}$. Referring to equations (56) and (69) and bearing in mind the above numerical values, we see that for the conducting layers we have

$$\begin{aligned}\kappa_1 &\approx \sigma_1 = (1+i)/\delta_1, \\ \eta_{1y} = \eta_{1\rho} &\approx \eta_1 = (1+i)/g_1\delta_1,\end{aligned}\tag{79}$$

to a very good approximation, since in our applications the quantity γ will always be of the order of $2\pi i/\lambda_v$, where the vacuum wavelength λ_v is at least a few meters, while the skin thickness δ_1 will be at most a small fraction of a centimeter.

For the insulating layers of thickness t_2 we shall set the conductivity g_2 equal to zero, so that

$$\sigma_2 = i\omega\sqrt{\mu_2\epsilon_2}, \quad \eta_2 = \sqrt{\mu_2/\epsilon_2}.\tag{80}$$

We denote the *relative* dielectric constant and permeability by ϵ_{2r} and μ_{2r} respectively; dissipation in the insulating layers may be included

if necessary by making ϵ_{2r} and/or μ_{2r} complex. In MKS units we have

$$\epsilon_2 = \epsilon_{2r}\epsilon_v, \quad \mu_2 = \mu_{2r}\mu_v, \quad (81)$$

where the electrical constants of vacuum are

$$\begin{aligned} \epsilon_v &= 8.854 \times 10^{-12} \text{ farads} \cdot \text{meter}^{-1}, \\ \mu_v &= 1.257 \times 10^{-6} \text{ henrys} \cdot \text{meter}^{-1}. \end{aligned} \quad (82)$$

It follows that

$$\begin{aligned} \sigma_2 &= \sigma_v \sqrt{\mu_{2r}\epsilon_{2r}} = \frac{2\pi i \sqrt{\mu_{2r}\epsilon_{2r}}}{\lambda_v} = \frac{2\pi i f_{\text{Mc}} \sqrt{\mu_{2r}\epsilon_{2r}}}{299.8} \text{ meters}^{-1}, \\ \eta_2 &= \eta_v \sqrt{\mu_{2r}/\epsilon_{2r}} = 376.7 \sqrt{\mu_{2r}/\epsilon_{2r}} \text{ ohms}, \end{aligned} \quad (83)$$

where as usual the subscript v refers to vacuum. It is clear that unless we deal with ferromagnetics, the quantities σ_2 and η_2 will be of roughly the same order of magnitude as σ_v and η_v . From (56) and (69) we have

$$\begin{aligned} \kappa_2 &= \sigma_2(1 - \gamma^2/\sigma_2^2)^{\frac{1}{2}}, \\ \eta_{2y} &= \eta_{2\rho} = \eta_2(1 - \gamma^2/\sigma_2^2)^{\frac{1}{2}}, \end{aligned} \quad (84)$$

where since σ_2 and γ are both of the same order of magnitude as $2\pi i/\lambda_v$, in general no further approximations can be made.

In all of what follows we shall assume that the thickness t_2 of each insulating layer is very small compared to the vacuum wavelength at the highest operating frequency; in practice t_2 will be at most a few mils and λ_v at least a few meters. Then the quantity $|\kappa_2 t_2|$, which is of the order of $2\pi t_2/\lambda_v$, will be so small that to an excellent approximation we may set $\text{sh } \kappa_2 t_2 = \kappa_2 t_2$ and $\text{ch } \kappa_2 t_2 = 1$. Using this simplification, together with the fact that $\eta_{1y} \ll \eta_{2y}$ for all frequencies which may conceivably be of interest, it is not difficult to show from (58) that the matrix elements of the plane double layer reduce to

$$\begin{aligned} \mathfrak{A} &= \text{ch } \kappa_1 t_1, \\ \mathfrak{B} &= \eta_{2y} \kappa_2 t_2 \text{ch } \kappa_1 t_1 + \eta_{1y} \text{sh } \kappa_1 t_1, \\ \mathfrak{C} &= \frac{1}{\eta_{1y}} \text{sh } \kappa_1 t_1, \\ \mathfrak{D} &= \frac{\eta_{2y} \kappa_2 t_2}{\eta_{1y}} \text{sh } \kappa_1 t_1 + \text{ch } \kappa_1 t_1. \end{aligned} \quad (85)$$

The determinant of the matrix is unity, and from (61) the propagation constant per section is defined by

$$\operatorname{ch} \Gamma = \frac{\eta_{2y}\kappa_2 t_2}{2\eta_{1y}} \operatorname{sh} \kappa_1 t_1 + \operatorname{ch} \kappa_1 t_1, \quad (86)$$

while from (63) the iterative impedances are

$$\begin{aligned} K_1 &= -\frac{1}{2}\eta_{2y}\kappa_2 t_2 + \sqrt{\left(\frac{1}{2}\eta_{2y}\kappa_2 t_2\right)^2 + \eta_{1y}\eta_{2y}\kappa_2 t_2 \coth \kappa_1 t_1 + \eta_{1y}^2}, \\ K_2 &= +\frac{1}{2}\eta_{2y}\kappa_2 t_2 + \sqrt{\left(\frac{1}{2}\eta_{2y}\kappa_2 t_2\right)^2 + \eta_{1y}\eta_{2y}\kappa_2 t_2 \coth \kappa_1 t_1 + \eta_{1y}^2}. \end{aligned} \quad (87)$$

If we make the same simplifications in the approximate expressions (73) for the matrix elements of a coaxial double layer, we obtain

$$\begin{aligned} \alpha &= \left[1 + \frac{t_1}{2\bar{\rho}}\right] \operatorname{ch} \kappa_1 t_1 - \frac{1}{2\kappa_1 \bar{\rho}} \operatorname{sh} \kappa_1 t_1, \\ \beta &= \left[1 + \frac{t_1 + t_2}{2\bar{\rho}}\right] \eta_{2\rho}\kappa_2 t_2 \operatorname{ch} \kappa_1 t_1 \\ &\quad + \left[1 + \frac{t_1}{2\bar{\rho}} + \left(2 - \frac{\eta_{2\rho}\kappa_2}{\eta_{1\rho}\kappa_1}\right) \frac{t_2}{2\bar{\rho}}\right] \eta_{1\rho} \operatorname{sh} \kappa_1 t_1, \\ \epsilon &= \left[1 + \frac{t_1}{2\bar{\rho}}\right] \frac{1}{\eta_{1\rho}} \operatorname{sh} \kappa_1 t_1, \\ \mathfrak{D} &= \left[1 + \frac{t_1 + t_2}{2\bar{\rho}} + \frac{\eta_{1\rho}}{2\eta_{2\rho}\kappa_1\kappa_2 t_2 \bar{\rho}}\right] \frac{\eta_{2\rho}\kappa_2 t_2}{\eta_{1\rho}} \operatorname{sh} \kappa_1 t_1 \\ &\quad + \left[1 + \frac{t_1 + 2t_2}{2\bar{\rho}}\right] \operatorname{ch} \kappa_1 t_1. \end{aligned} \quad (88)$$

In the preceding equations no restrictions have been laid on the thicknesses t_1 and t_2 except the trivial requirement that t_2 shall be small compared to a wavelength. We shall now consider the limiting case in which both t_1 and t_2 are infinitesimally small. When we make this last and most drastic approximation we do not expect that the idealized structure thus obtained will show all of the features which are of interest in a physical transmission line with finite layers; but the results of the simplified analysis will be useful in some cases nevertheless. It need scarcely be pointed out that we are dealing here only with a mathematical limiting process, in which we assume that each layer, no matter how thin, always exhibits the same electrical properties as the bulk material. If this assumption be regarded as unrealistic, it may be observed that the quantity which we actually allow to tend to zero is the ratio of layer thickness to skin depth. The skin depth may be made as large as desired by lowering the frequency, so that the formulas which we derive by

letting t_1 and t_2 approach zero at a finite frequency will also hold for finite thicknesses if the frequency is sufficiently low.

We shall let θ denote the fraction of the stack which is occupied by conducting material, so that

$$\theta = t_1/(t_1 + t_2), \quad (89)$$

where at present t_1 and t_2 are both infinitesimal. Then the stack may be regarded as a homogeneous, anisotropic medium, characterized by an average dielectric constant $\bar{\epsilon}$ perpendicular to the layers, an average permeability $\bar{\mu}$ parallel to the layers, and an average conductivity \bar{g} parallel to the layers. Sakurai¹¹ has treated such an artificial anisotropic medium, and from his formulas we find that when the layers are alternately conductors and insulators, the average electrical constants are, to a very good approximation,

$$\begin{aligned} \bar{\epsilon} &= \epsilon_2/(1 - \theta), \\ \bar{\mu} &= \theta\mu_1 + (1 - \theta)\mu_2, \\ \bar{g} &= \theta g_1. \end{aligned} \quad (90)$$

Sakurai has also shown that the average values of the electrical constants may be used in Maxwell's equations for the average (macroscopic) fields, due regard being paid to the orientations of the field vectors with respect to the laminae.

For the plane stack, these equations read

$$\begin{aligned} \partial \bar{H}_x / \partial z &= i\omega \bar{\epsilon} \bar{E}_y, \\ \partial \bar{H}_x / \partial y &= -\bar{g} \bar{E}_z, \\ \partial \bar{E}_y / \partial z - \partial \bar{E}_z / \partial y &= i\omega \bar{\mu} \bar{H}_x, \end{aligned} \quad (91)$$

where the bars denote average values. By analysis exactly similar to that carried out at the beginning of this section for a homogeneous, isotropic medium, we may find the relation between the tangential field components E_z , H_x at the two surfaces of a stack of infinitesimally thin layers. (The bars representing average values may be omitted, since the tangential components of \mathbf{E} and \mathbf{H} are continuous across the boundaries of the layers.) We obtain a matrix relation analogous to (55), namely

$$\begin{pmatrix} E(0) \\ H(0) \end{pmatrix} = \begin{pmatrix} \text{ch } \Gamma_\epsilon s & K \text{ sh } \Gamma_\epsilon s \\ \frac{1}{K} \text{ sh } \Gamma_\epsilon s & \text{ch } \Gamma_\epsilon s \end{pmatrix} \begin{pmatrix} E(s) \\ H(s) \end{pmatrix}, \quad (92)$$

¹¹ T. Sakurai, *J. Phys. Soc. Japan*, **5**, 394 (1950), especially Section 3.

where s is the thickness of the stack. The propagation constant Γ_ℓ per unit distance normal to the stack and the characteristic impedance K of the stack are given by

$$\Gamma_\ell = \left[\frac{i\bar{g}}{\omega\bar{\epsilon}} (\omega^2\bar{\mu}\bar{\epsilon} + \gamma^2) \right]^{\frac{1}{2}}, \quad (93)$$

$$K = \Gamma_\ell / \bar{g} = \left[\frac{i}{\omega\bar{\epsilon}\bar{g}} (\omega^2\bar{\mu}\bar{\epsilon} + \gamma^2) \right]^{\frac{1}{2}}. \quad (94)$$

Γ_ℓ and K may also be derived from equations (86) and (87) by limiting processes; we have

$$\Gamma_\ell = \lim_{t_1+t_2 \rightarrow 0} \Gamma / (t_1 + t_2), \quad (95)$$

$$K = \lim_{t_1+t_2 \rightarrow 0} K_1 = \lim_{t_1+t_2 \rightarrow 0} K_2. \quad (96)$$

It should perhaps be noted that terms of the order of $\omega\epsilon_1/g_1$ and $\omega\epsilon_2/g_1$ compared to unity were omitted in the expressions (90) for $\bar{\epsilon}$ and \bar{g} , and in the derivations of Γ_ℓ and K . Since, however, under all practical circumstances the omitted terms appear to be insignificant, we shall not take space to write out the formally more complicated results which would be obtained by keeping them.¹²

In a cylindrical stack of infinitesimal layers, the average fields satisfy

$$\begin{aligned} \partial \bar{H}_\phi / \partial z &= -i\omega\bar{\epsilon}\bar{E}_\rho, \\ \partial(\rho\bar{H}_\phi) / \partial \rho &= \bar{g}\rho\bar{E}_z, \\ \partial\bar{E}_z / \partial \rho - \partial\bar{E}_\rho / \partial z &= i\omega\bar{\mu}\bar{H}_\phi. \end{aligned} \quad (97)$$

The relation between the tangential field components E_z , $-H_\phi$ at two radii ρ_0 and ρ_n is expressed by a matrix equation analogous to (68), namely

$$\begin{pmatrix} E(\rho_0) \\ -H(\rho_0) \end{pmatrix} = \begin{pmatrix} \Gamma_\ell \rho_n (K_{00} I_{1n} + K_{1n} I_{00}) & K \Gamma_\ell \rho_n (K_{00} I_{0n} - K_{0n} I_{00}) \\ \frac{\Gamma_\ell \rho_n}{K} (K_{10} I_{1n} - K_{1n} I_{10}) & \Gamma_\ell \rho_n (K_{10} I_{0n} + K_{0n} I_{10}) \end{pmatrix} \begin{pmatrix} E(\rho_n) \\ -H(\rho_n) \end{pmatrix}, \quad (98)$$

¹² In Reference 1, equations (II-17) through (II-26) give examples of equations in which these small terms have been retained.

where

$$I_{rs} = I_r(\Gamma_t \rho_s), \quad K_{rs} = K_r(\Gamma_t \rho_s), \quad (99)$$

and Γ_t and K are given, as in the plane case, by (93) and (94).

IV. PRINCIPAL MODE IN CLOGSTON 1 LINES WITH INFINITESIMALLY THIN LAMINAE

An idealized parallel-plane Clogston 1 transmission line is shown schematically in Fig. 5. It consists of a slab of dielectric of thickness b , with electrical constants μ_0 , ϵ_0 , bounded above and below by laminated stacks each of thickness s . Outside each stack there may be an insulating or a conducting sheath, of which nothing more will be assumed at present than that its normal surface impedance $Z_n(\gamma)$ is known. The total distance between the sheaths will be denoted by a , where $a = b + 2s$.

The corresponding Clogston 1 coaxial line is shown in Fig. 6. We denote the thickness of the inner and outer stacks by s_1 and s_2 respectively, while a is the radius of the inner core (if any), and b is the inner radius of the sheath around the outer stack. The inner and outer radii of the main dielectric are $\rho_1 = a + s_1$ and $\rho_2 = b - s_2$, respectively. In practice the core may be a dielectric rod and the sheath may be a conducting shield, but in the present theoretical analysis we shall merely assume that the radial impedances $Z_a(\gamma)$ and $Z_b(\gamma)$ looking into the core and the sheath are known.

In Part I of this paper we shall deal with "extreme" Clogston 1 lines, in which the space occupied by the stacks is small compared to the space occupied by the main dielectric. We may then regard the laminated boundaries as impedance sheets guiding waves whose phase velocity is

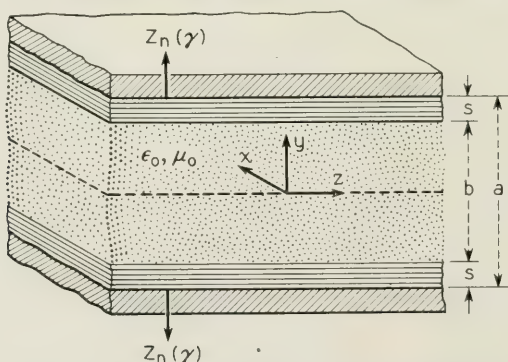


Fig. 5—Parallel-plane Clogston 1 transmission line.

determined by the properties of the main dielectric, as discussed in Section II, and we may use the intrinsic propagation constant of the main dielectric in calculating the surface impedance of the boundaries. This approximation simplifies the analysis of Clogston 1 lines a great deal. We shall treat the general case, in which an arbitrary fraction of the total space is filled with laminations, in Section IX of Part II, as a part of our study of Clogston 2 lines.

In this section we shall assume that the laminae are infinitesimally thin, so that the stacks may be completely characterized by their average properties $\bar{\epsilon}$, $\bar{\mu}$, and \bar{g} . The case of finite laminae will be taken up in the next section. We shall also assume throughout that dielectric and magnetic dissipation may be neglected except, as in Section VII, where the contrary is explicitly stated.

In general the current density and the other field quantities in a plane stack of infinitesimally thin layers will be linear combinations of the functions $\text{sh } \Gamma_\ell y$ and $\text{ch } \Gamma_\ell y$, where y is distance measured into the stack, and Γ_ℓ is the propagation constant per unit distance, as given by (93). The qualitative behavior of the fields in a cylindrical stack will be similar. In particular, if the stack is thick enough the current density and the fields will fall off as $e^{-\Gamma_\ell y}$, and we can define an "effective skin depth" Δ by

$$\Delta = 1/(\text{Re } \Gamma_\ell). \quad (100)$$

Clogston's fundamental observation was that in order to minimize the

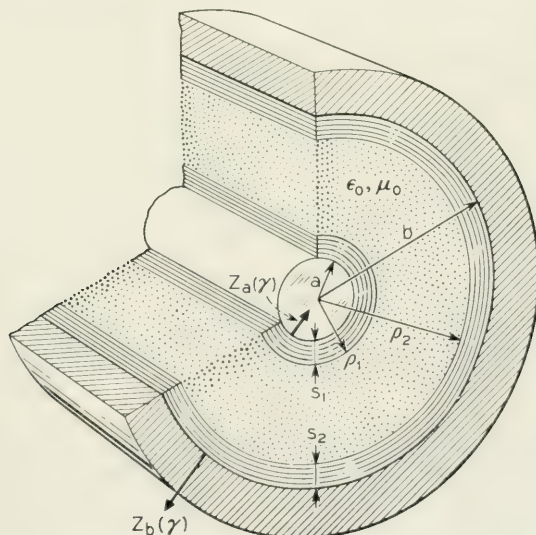


Fig. 6—Coaxial Clogston 1 transmission line.

ohmic losses in a stack carrying a fixed total current the current density should be uniform across the stack, and that we can achieve uniform current density by adjusting the $\mu_0\epsilon_0$ product of the main dielectric so as to make Γ_t equal to zero. If in equation (93) we set

$$\gamma = \gamma_0 = i\omega\sqrt{\mu_0\epsilon_0}, \quad (101)$$

then Γ_t will be zero if

$$\mu_0\epsilon_0 = \bar{\mu}\bar{\epsilon} = [\theta\mu_1 + (1 - \theta)\mu_2][\epsilon_2/(1 - \theta)]. \quad (102)$$

Equation (102) will be referred to henceforth as *Clogston's condition*.¹³ If the permeabilities of the various materials are all equal, the condition reduces to

$$\epsilon_0 = \bar{\epsilon} = \epsilon_2/(1 - \theta), \quad (103)$$

which is the form employed by Clogston in Reference 1.

When Clogston's condition is satisfied, $\Gamma_t = 0$ and the effective skin depth of the stack is infinite;¹³ that is, the current density is uniform in any stack of finite total thickness. The quantities Γ_t and K vanish simultaneously, but the limiting value of their ratio is finite; and the matrix of the plane stack, as given by (92), takes the form

$$\begin{pmatrix} 1 & 0 \\ \bar{g}s & 1 \end{pmatrix}. \quad (104)$$

Accordingly we obtain, for the surface impedance $Z_0(\gamma_0)$ of the stack,

$$Z_0(\gamma_0) = \frac{1}{\bar{g}s + 1/Z_n(\gamma_0)}, \quad (105)$$

which is, as might have been expected, just the impedance between opposite edges of a unit square of material of conductivity \bar{g} and thickness s through which the current density is uniform, in parallel with the sheath impedance $Z_n(\gamma_0)$. It follows from equations (20) and (21) of Section II that the attenuation and phase constants of the principal mode in a plane Clogston 1 line with infinitesimally thin laminae, Clogston's condition being satisfied exactly, are

¹³ This statement is certainly accurate enough for all practical purposes, although an exact calculation which takes into account the small terms that were neglected in the approximate formula (93) for Γ_t shows that the effective skin depth is $\lambda_0/2\pi\theta$, where λ_0 is the length of a free wave in the main dielectric. The exact result is derived by Clogston in Reference 1, equation (II-26). In practice, finite lamina thickness will restrict us to effective skin depths much smaller than this theoretical limit.

$$\alpha = \operatorname{Re} \frac{1}{\eta_0 b [\bar{g}s + 1/Z_n(\gamma_0)]}, \quad (106)$$

$$\beta = \omega \sqrt{\mu_0 \epsilon_0} + \operatorname{Im} \frac{1}{\eta_0 b [\bar{g}s + 1/Z_n(\gamma_0)]}. \quad (107)$$

In general the sheath impedance $Z_n(\gamma_0)$ will be large compared to the impedance $1/\bar{g}s$ of the stack, since even if the sheath is an electrically thick metal plate of the same material as the conducting layers, its impedance is

$$Z_n(\gamma_0) = (1 + i)/g_1 \delta_1, \quad (108)$$

whereas θs will usually be several times the skin thickness δ_1 in the frequency range of interest. If the sheath is free space, its impedance is a fortiori much greater than $1/\bar{g}s$, since then it may be shown that

$$Z_n(\gamma_0) = -i\eta_v(\mu_{0r}\epsilon_{0r} - 1)^{\frac{1}{2}}, \quad (109)$$

where $\eta_v = 376.7$ ohms is the intrinsic impedance of free space, and μ_{0r} and ϵ_{0r} are the relative permeability and relative dielectric constant of the main dielectric. Under most circumstances, therefore, we may neglect $1/Z_n(\gamma_0)$ in comparison with $\bar{g}s$, and obtain the very simple results,

$$\alpha = 1/\eta_0 b \bar{g}s, \quad (110)$$

$$\beta = \omega \sqrt{\mu_0 \epsilon_0}. \quad (111)$$

To this approximation the line exhibits neither amplitude nor phase distortion.

For a coaxial stack of infinitesimally thin layers with Clogston's condition satisfied, the stack matrix given in (98) reduces to

$$\begin{pmatrix} 1 & 0 \\ \frac{\bar{g}}{2\rho_0} (\rho_n^2 - \rho_0^2) & \frac{\rho_n}{\rho_0} \end{pmatrix}, \quad (112)$$

where ρ_0 and ρ_n denote the inner and outer radii of the stack. It follows from (112) that

$$\begin{aligned} \frac{Z_1(\gamma_0)}{\rho_1} &= \frac{1}{\frac{1}{2}\bar{g}(\rho_1^2 - a^2) + a/Z_a(\gamma_0)} = \frac{1}{\bar{g}s_1(a + \frac{1}{2}s_1) + a/Z_a(\gamma_0)}, \\ \frac{Z_2(\gamma_0)}{\rho_2} &= \frac{1}{\frac{1}{2}\bar{g}(b^2 - \rho_2^2) + b/Z_b(\gamma_0)} = \frac{1}{\bar{g}s_2(b - \frac{1}{2}s_2) + b/Z_b(\gamma_0)}, \end{aligned} \quad (113)$$

where $Z_1(\gamma_0)$ and $Z_2(\gamma_0)$ are the radial impedances looking into the stacks at ρ_1 and ρ_2 respectively, and $Z_a(\gamma_0)$ and $Z_b(\gamma_0)$ are the radial impedances looking into the core and the outer sheath. From equations (44) and (45) of Section II, the attenuation and phase constants of a coaxial Clogston 1 cable with infinitesimally thin layers, Clogston's condition being satisfied exactly, are

$$\alpha = \operatorname{Re} \frac{Z_1(\gamma_0)/\rho_1 + Z_2(\gamma_0)/\rho_2}{2\eta_0 \log (\rho_2/\rho_1)}, \quad (114)$$

$$\beta = \omega \sqrt{\mu_0 \epsilon_0} + \operatorname{Im} \frac{Z_1(\gamma_0)/\rho_1 + Z_2(\gamma_0)/\rho_2}{2\eta_0 \log (\rho_2/\rho_1)}, \quad (115)$$

where $Z_1(\gamma_0)/\rho_1$ and $Z_2(\gamma_0)/\rho_2$ are given by (113).

The impedances $Z_a(\gamma_0)$ and $Z_b(\gamma_0)$ may be computed if we know the structure of the core and the sheath. For a solid, homogeneous core and a homogeneous sheath of effectively infinite thickness, we have

$$Z_a(\gamma_0) = \frac{\eta \kappa}{\sigma} \frac{I_0(\kappa a)}{I_1(\kappa a)}, \quad Z_b(\gamma_0) = \frac{\eta \kappa}{\sigma} \frac{K_0(\kappa b)}{K_1(\kappa b)}, \quad (116)$$

where

$$\kappa = \sqrt{\sigma^2 - \gamma_0^2}, \quad (117)$$

but of course the intrinsic propagation constant σ and the intrinsic impedance η need not be the same for the core and the sheath. If the sheath is of finite electrical thickness or has a laminated structure (alternate layers of copper and iron, for example, to provide effective shielding), its surface impedance may be calculated by a straightforward but longer procedure. We shall not go into this matter here, but shall merely observe that in many cases of interest $Z_a(\gamma_0)$ and $Z_b(\gamma_0)$ are so large that we may neglect the terms containing their reciprocals in (113). This means that we neglect the total conduction and displacement currents flowing in the core and the sheath, compared to the conduction currents in the stacks. Then the expressions for the attenuation and phase constants become

$$\alpha = \frac{1}{2\eta_0 \bar{g} \log (\rho_2/\rho_1)} \left[\frac{1}{s_1(a + \frac{1}{2}s_1)} + \frac{1}{s_2(b - \frac{1}{2}s_2)} \right], \quad (118)$$

$$\beta = \omega \sqrt{\mu_0 \epsilon_0}, \quad (119)$$

and again to this approximation there is neither amplitude nor phase distortion.

The formulas which have just been derived on the assumption of

infinitesimally thin laminae approach validity for laminae of finite thickness as the frequency is reduced, provided of course that we do not go to such extremely low frequencies that the attenuation per wavelength becomes large. We shall show in the next section that the effect of finite lamina thickness is to introduce a frequency dependence into the attenuation and phase constants, in addition to the variations (if any) which arise from the frequency dependence of the core and sheath impedances.

We next write down approximate expressions for the field components in a plane Clogston 1 line with infinitesimally thin laminae. In the main dielectric we have, from equations (22) of Section II,

$$\begin{aligned} H_x &\approx H_0 e^{-\gamma z}, \\ E_y &\approx -\sqrt{\frac{\mu_0}{\epsilon_0}} H_0 e^{-\gamma z}, \\ E_z &\approx \frac{2Z_0(\gamma_0)H_0 y}{b} e^{-\gamma z}, \end{aligned} \quad (120)$$

for $-\frac{1}{2}b \leq y \leq \frac{1}{2}b$, where H_0 is an arbitrary amplitude factor and $Z_0(\gamma_0)$ is given by (105). In the stacks the fields are

$$\begin{aligned} H_x &\approx H_0[1 + \bar{g}Z_0(\gamma_0)(\tfrac{1}{2}b \mp y)]e^{-\gamma z}, \\ \bar{E}_y &\approx -\sqrt{\frac{\bar{\mu}}{\bar{\epsilon}}} H_0[1 + \bar{g}Z_0(\gamma_0)(\tfrac{1}{2}b \mp y)]e^{-\gamma z}, \\ E_z &\approx \pm Z_0(\gamma_0)H_0 e^{-\gamma z}, \end{aligned} \quad (121)$$

for $\frac{1}{2}b \leq |y| \leq \frac{1}{2}a$, where in cases of ambiguous sign the upper sign refers to the upper stack ($y > 0$) and the lower sign to the lower stack ($y < 0$). It should be noted that whereas the tangential field components H_x and E_z are continuous through the stack, the normal field component E_y is discontinuous at layer boundaries. From equation (52) we have, in the conducting layers,

$$E_y = -(\gamma/g_1)H_x, \quad (122)$$

while in the insulating layers,

$$E_y = -(\gamma/i\omega\epsilon_2)H_x. \quad (123)$$

To our approximation, therefore, the only contributions to the average field \bar{E}_y come from the insulating layers.

The average current density \bar{J}_z in either stack is uniform, being

given by

$$\bar{J}_z = \bar{g}E_z = \pm \bar{g}Z_0(\gamma_0)H_0e^{-\gamma z}. \quad (124)$$

The total current per unit width carried by the stack is just \bar{J}_zs , where s is the thickness of the stack; there will also be small currents in the sheaths unless we assume the sheath impedance to be infinite. The potential difference between any two points y_1 and y_2 in the same transverse plane may easily be found from

$$V(y_2) - V(y_1) = - \int_{y_1}^{y_2} E_y dy. \quad (125)$$

For a Clogston 1 line of the proportions which we have been considering, the potential difference across the stacks will be small compared to the potential difference across the main dielectric.

In a coaxial Clogston 1 with infinitesimally thin laminae, the fields in the main dielectric are given to a good approximation by equations (46) of Section II, namely

$$\begin{aligned} H_\phi &\approx \frac{I}{2\pi\rho} e^{-\gamma z}, \\ E_\rho &\approx \sqrt{\frac{\mu_0}{\epsilon_0}} \frac{I}{2\pi\rho} e^{-\gamma z}, \\ E_z &\approx \frac{I}{2\pi \log(\rho_2/\rho_1)} \left[\frac{Z_1(\gamma_0)}{\rho_1} \log \frac{\rho_2}{\rho} + \frac{Z_2(\gamma_0)}{\rho_2} \log \frac{\rho_1}{\rho} \right] e^{-\gamma z}, \end{aligned} \quad (126)$$

where I is an arbitrary amplitude factor and $Z_1(\gamma_0)$ and $Z_2(\gamma_0)$ are expressed by (113). In the inner stack we have

$$\begin{aligned} H_\phi &\approx \frac{Z_1(\gamma_0)I}{2\pi\rho_1} \left[\frac{\bar{g}(\rho^2 - a^2)}{2\rho} + \frac{a}{\rho Z_a(\gamma_0)} \right] e^{-\gamma z}, \\ \bar{E}_\rho &\approx \sqrt{\frac{\bar{\mu}}{\bar{\epsilon}}} \frac{Z_1(\gamma_0)I}{2\pi\rho_1} \left[\frac{\bar{g}(\rho^2 - a^2)}{2\rho} + \frac{a}{\rho Z_a(\gamma_0)} \right] e^{-\gamma z}, \\ E_z &\approx \frac{Z_1(\gamma_0)I}{2\pi\rho_1} e^{-\gamma z}, \end{aligned} \quad (127)$$

while in the outer stack,

$$\begin{aligned}
H_\phi &\approx \frac{Z_2(\gamma_0)I}{2\pi\rho_2} \left[\frac{\bar{g}(b^2 - \rho^2)}{2\rho} + \frac{b}{\rho Z_b(\gamma_0)} \right] e^{-\gamma z}, \\
\bar{E}_\rho &\approx \sqrt{\frac{\bar{\mu}}{\bar{\epsilon}}} \frac{Z_2(\gamma_0)I}{2\pi\rho_2} \left[\frac{\bar{g}(b^2 - \rho^2)}{2\rho} + \frac{b}{\rho Z_b(\gamma_0)} \right] e^{-\gamma z}, \\
E_z &\approx -\frac{Z_2(\gamma_0)I}{2\pi\rho_2} e^{-\gamma z}.
\end{aligned} \tag{128}$$

The average current density in either stack is uniform and is given by

$$\bar{J}_z = \bar{g}E_z, \tag{129}$$

though in general the current density will not be the same in the two stacks because of the difference in cross-sectional areas. The potential difference between the surface of the inner core and any other point in the same transverse plane is

$$V(\rho) - V(a) = -\int_a^\rho E_\rho d\rho. \tag{130}$$

If the stacks are thin compared to the thickness of the main dielectric, as we are assuming throughout Part I, then the potential difference across the stacks will be small compared to the potential difference across the main dielectric, and the characteristic impedance Z_k of the Clogston 1 cable will be approximately the same as the characteristic impedance of an ideal coaxial cable with perfect conductors of radii ρ_1 and ρ_2 and the same main dielectric, namely

$$Z_k = 60 \sqrt{\frac{\mu_{0r}}{\epsilon_{0r}}} \log \frac{\rho_2}{\rho_1} \text{ ohms.} \tag{131}$$

We shall defer making any field plots for Clogston-type transmission lines until Section IX of Part II, when we shall discuss the transition from Clogston 1 to Clogston 2 as the space originally occupied by the main dielectric is gradually filled with laminations. Our present results will then appear as the limiting case in which the thickness of the stacks is small compared to the thickness of the main dielectric.

In conclusion we shall mention briefly the question of how to dispose a given amount of laminated material in a Clogston 1 coaxial cable so as to achieve the minimum attenuation constant. The whole problem of optimum proportions for Clogston cables is a complicated one of which an adequate treatment would require a separate paper in itself, with the results depending to a great extent on engineering considerations which limit the ranges of the parameters that we can vary in any practical case. Here we shall discuss only the following rather highly idealized problem:

Given a coaxial Clogston 1 with infinitesimally thin laminae, having a high-impedance core and a high-impedance sheath of fixed radius b , and in which the total thickness $s_1 + s_2$ of both stacks is a fixed constant $2s$. Assuming that $2s$ is small compared to b , what should be the radius a of the core, and how should the total stack thickness be divided between the outer and inner stacks so as to minimize the attenuation constant of the line? Finally, what should be the fraction θ of conducting material in the stacks to minimize the attenuation constant, if the electrical constants of the conducting and insulating layers are fixed, but the properties of the main dielectric are at our disposal?

If the two inequalities

$$s_1 \ll a, \quad s_2 \ll b, \quad (132)$$

are satisfied (these restrictions will be removed in Section IX, when we discuss Clogston cables having an arbitrary fraction of their total volume filled with laminations), then equation (118) for the attenuation constant of a Clogston 1 with infinitesimally thin laminae and high-impedance boundaries becomes, approximately,

$$\alpha \approx \frac{1}{2\eta_0 \bar{g} \log(b/a)} \left[\frac{1}{as_1} + \frac{1}{bs_2} \right]. \quad (133)$$

If we write

$$s_2 = 2s - s_1, \quad (134)$$

and vary s_1 and s_2 in accordance with this relation while holding a and b constant, it is easy to show that the expression on the right side of (133) is a minimum when

$$s_1 = \frac{2s\sqrt{b}}{\sqrt{a} + \sqrt{b}}, \quad s_2 = \frac{2s\sqrt{a}}{\sqrt{a} + \sqrt{b}}. \quad (135)$$

These equations tell us the most efficient way to divide the stacks in a Clogston 1 when the radii of the core and the outer sheath are a and b respectively, still assuming of course that the thickness of each stack is small compared to its mean radius.

If we introduce the optimum values of s_1 and s_2 into (133), we get

$$\alpha \approx \frac{1}{2\eta_0 \bar{g}(s_1 + s_2) \log(b/a)} \left[\frac{1}{\sqrt{a}} + \frac{1}{\sqrt{b}} \right]^2. \quad (136)$$

If b is fixed, the last expression is a minimum, considered as a function of a , when

$$\log(b/a) = 1 + \sqrt{a/b}. \quad (137)$$

The root of this transcendental equation is

$$b/a = 4.3827, \quad a = 0.22817b. \quad (138)$$

Substituting this value of b/a into (135), we find

$$\begin{aligned} s_1 &= 1.3535s, \\ s_2 &= 0.6465s, \\ s_1/s_2 &= 2.0935; \end{aligned} \quad (139)$$

while from (136) and (138) the minimum value of the attenuation constant is

$$\alpha \approx \frac{3.238}{\eta_0 \bar{g}(s_1 + s_2)b}. \quad (140)$$

To find the optimum value of θ , we observe that equation (118) for the attenuation constant of a Clogston 1 cable with infinitesimally thin laminae and high-impedance boundaries may be written in the form

$$\alpha = \frac{(\epsilon_0/\mu_0)^{\frac{1}{2}}}{\theta g_1} f(a, b, s_1, s_2), \quad (141)$$

where the first factor depends on the electrical constants of the components of the cable, while $f(a, b, s_1, s_2)$ is a function only of the geometry. By (110) the attenuation constant of a plane Clogston 1 has the same form, only with a different dependence on the geometrical factors. Now assume that the geometrical proportions of the line are fixed, and that the electrical constants μ_1, g_1, μ_2 , and ϵ_2 of the conducting and insulating layers are given, but that the constants μ_0, ϵ_0 of the main dielectric and the fraction of space θ occupied by conducting layers in the stacks are at our disposal. The $\mu_0 \epsilon_0$ product of the main dielectric is to be codetermined with θ so that Clogston's condition (102) is always satisfied. Solving (102) for θ gives

$$\theta = \frac{\mu_0 \epsilon_0 - \mu_2 \epsilon_2}{\mu_0 \epsilon_0 + (\mu_1 - \mu_2) \epsilon_2}. \quad (142)$$

Hence the first factor in the expression (141) for α may be written

$$\frac{(\epsilon_0/\mu_0)^{\frac{1}{2}}}{\theta g_1} = \frac{\epsilon_0^{\frac{1}{2}} [\mu_0 \epsilon_0 + (\mu_1 - \mu_2) \epsilon_2]}{g_1 \mu_0^{\frac{1}{2}} [\mu_0 \epsilon_0 - \mu_2 \epsilon_2]}. \quad (143)$$

If we minimize the right side of (143) with respect to ϵ_0 , all other quantities being held constant, by equating to zero the derivative with respect

to ϵ_0 and then solving for ϵ_0 , we get

$$\mu_0 \epsilon_0 = \frac{1}{2}[(\mu_1 + 2\mu_2) + (\mu_1^2 + 8\mu_1\mu_2)^{\frac{1}{2}}]\epsilon_2. \quad (144)$$

From (142) the value of θ is

$$\theta = \frac{\mu_1 + (\mu_1^2 + 8\mu_1\mu_2)^{\frac{1}{2}}}{3\mu_1 + (\mu_1^2 + 8\mu_1\mu_2)^{\frac{1}{2}}}, \quad (145)$$

and the corresponding attenuation constant is proportional to

$$\frac{(\epsilon_0/\mu_0)^{\frac{1}{2}}}{\theta g_1} = \frac{(\epsilon_0/\mu_0)^{\frac{1}{2}}}{g_1} \frac{3\mu_1 + (\mu_1^2 + 8\mu_1\mu_2)^{\frac{1}{2}}}{\mu_1 + (\mu_1^2 + 8\mu_1\mu_2)^{\frac{1}{2}}}. \quad (146)$$

It will be observed that so far we have determined only the optimum value of the product $\mu_0 \epsilon_0$, and so we are still free to alter the ratio of μ_0 to ϵ_0 while holding the product of these two quantities constant. For given values of μ_1 and μ_2 , we obtain the lowest attenuation constant by making ϵ_0 as small as possible and μ_0 as large as possible, subject of course to the practical restriction that ϵ_0 cannot be lower than the dielectric constant of free space. However if we permit μ_2 and μ_0 to be simultaneously increased, as by magnetic loading of both the insulating layers and the main dielectric, we find from (146) that on paper it is possible to decrease the attenuation constant without any definite limit. This observation is in accord with the fact that the attenuation constant of an ordinary coaxial cable may be decreased indefinitely, with a corresponding decrease in the velocity of propagation along the cable, if we are willing to assume an unlimited amount of lossless magnetic loading.

If $\mu_1 = \mu_2$, (144) and (145) take the form

$$\mu_0 \epsilon_0 = 3\mu_2 \epsilon_2, \quad \theta = 2/3, \quad (147)$$

from which we have the result given by Clogston:¹⁴ If the conducting and insulating layers are infinitesimally thin and have equal permeabilities, then minimum attenuation is achieved when *the thickness of the conducting layers is twice the thickness of the insulating layers*. In this case, from (146) and (147) the attenuation is proportional to

$$\frac{(\epsilon_0/\mu_0)^{\frac{1}{2}}}{\theta g_1} = \frac{3(\epsilon_0/\mu_0)^{\frac{1}{2}}}{2g_1}. \quad (148)$$

When $\mu_0 = \mu_2$, corresponding to no magnetic loading, we must take $\epsilon_0 = 3\epsilon_2$, and (148) reduces to

¹⁴ Reference 1, pp. 513-514.

$$\frac{(\epsilon_0/\mu_0)^{\frac{1}{2}}}{\theta g_1} = \frac{3\sqrt{3} (\epsilon_2/\mu_2)^{\frac{1}{2}}}{2g_1}, \quad (149)$$

while if we load the main dielectric so that $\mu_0 = 3\mu_2$ and we can take $\epsilon_0 = \epsilon_2$, we have

$$\frac{(\epsilon_0/\mu_0)^{\frac{1}{2}}}{\theta g_1} = \frac{\sqrt{3} (\epsilon_2/\mu_2)^{\frac{1}{2}}}{2g_1}, \quad (150)$$

which is just one-third of the value with no magnetic loading.

As Clogston has pointed out, if the limitation is on the total thickness of conducting material in the stacks rather than on the stack thicknesses themselves, we shall find it advantageous to use a small value of θ (a high "dilution" of conducting material) so as to make the average dielectric constant $\epsilon_2/(1 - \theta)$ of the stacks, which has to be matched by the main dielectric, as small as possible. We shall see later that the effect of finite lamina thickness is in fact to limit the total thickness of conducting material which it is useful to employ in a single stack at high frequencies, so that for physical stacks of non-magnetic layers at high frequencies the optimum value of θ is less than $2/3$. Quantitative results which take into account the finite thickness of the layers will be obtained in Section XI.

To illustrate the use of some of the equations derived above by means of a numerical example, we shall compare the attenuation constant of a conventional coaxial cable with that of a Clogston 1 cable of the same size. If a and b denote the radii of the inner and outer conductors of a conventional coaxial cable, and we take $b/a = 3.5911$ to minimize the attenuation constant, then we have from equation (48) of Section II, on setting $\rho_1 = a$ and $\rho_2 = b$,

$$\alpha = \frac{1.796}{\eta_0 g_1 \delta_1 b}, \quad (151)$$

where η_0 is the intrinsic impedance of the main dielectric, which may be air. For a Clogston 1 coaxial cable with infinitesimally thin laminae, no magnetic material in the stacks ($\mu_1 = \mu_2 = \mu_v$), and the optimum proportions given by (139) and (147), we have

$$\alpha \approx \frac{4.857}{\eta_0 g_1 (s_1 + s_2) b}, \quad (152)$$

where b is the outside radius of the outer stack and η_0 is the intrinsic impedance of the main dielectric, which cannot be air in a Clogston cable. The ratio of the attenuation constant α_c of this Clogston cable to the

attenuation constant α_s of an *air-filled* standard coaxial of the same size, made of the same conducting material, is

$$\frac{\alpha_c}{\alpha_s} \approx \frac{2.704 \delta_1}{(\mu_{0r}/\epsilon_{0r})^{1/2} (s_1 + s_2)}, \quad (153)$$

where μ_{0r} and ϵ_{0r} refer to the main dielectric of the Clogston cable.

Since the attenuation constant of a standard coaxial cable is proportional to the square root of frequency in the range we are considering, while the attenuation constant of the ideal Clogston cable is independent of frequency in this range, there will be a crossover frequency above which the Clogston cable has a lower attenuation constant than a conventional coaxial cable of the same size. If we are dealing with copper conductors and if frequencies are measured in $\text{Mc} \cdot \text{sec}^{-1}$ and linear dimensions in mils, then from equations (78) and (153) we find that the crossover frequency is given approximately by

$$f_{\text{Mc}} \approx \frac{49.50 (\epsilon_{0r}/\mu_{0r})}{(s_1 + s_2)_{\text{mils}}^2}. \quad (154)$$

For example, let us take an ideal Clogston 1 cable of outer diameter 0.375 inches, excluding the sheath, with no magnetic loading, and assume the following values:

$$\begin{aligned} a &= 42.8 \text{ mils} & \theta &= 2/3 \\ b &= 187.5 \text{ mils} & \epsilon_{2r} &= 2.26 \text{ (polyethylene)} \\ s_1 &= 12.69 \text{ mils} & \epsilon_{0r} &= 3\epsilon_{2r} = 6.78 \\ s_2 &= 6.06 \text{ mils} & \mu_{0r} &= \mu_{1r} = \mu_{2r} = 1 \\ s_1 + s_2 &= 18.75 \text{ mils} \end{aligned} \quad (155)$$

This cable has a lower attenuation constant than a standard air-filled coaxial of the same size at frequencies above about $1 \text{ Mc} \cdot \text{sec}^{-1}$, the approximate formula (154) yielding $0.955 \text{ Mc} \cdot \text{sec}^{-1}$ for the crossover frequency and the exact equation (118), taken in conjunction with (151), yielding $1.251 \text{ Mc} \cdot \text{sec}^{-1}$.

The reader is cautioned that the comparison given by (153) between Clogston and conventional cables is based upon certain highly idealized assumptions. In the first place we have neglected the finite thickness of the laminae, which will in fact cause the attenuation constant of a physical Clogston cable to increase with increasing frequency, and ultimately to cross over again and become higher than the attenuation constant of a conventional air-filled coaxial. We have also neglected dielectric and magnetic losses, which are likely to be directly proportional to frequency and by no means negligible at the upper end of the

frequency band. In practice, too, the $\mu_0\epsilon_0$ product of the main dielectric must be held very close to the Clogston value or the benefit of the large effective skin depth is lost; and the stacks must be extremely uniform or again the depth of penetration is greatly reduced. We shall take up all these matters in later sections, and shall see that while the results just given represent ultimate limits of performance, the practical improvements which can be achieved over conventional cables depend upon the degree to which one can solve the manufacturing problems that tend to make every actual Clogston cable differ more or less from the ideal structure considered above.

V. EFFECT OF FINITE LAMINA THICKNESS. FREQUENCY DEPENDENCE OF ATTENUATION IN CLOGSTON 1 LINES

The principal effect of finite lamina thickness in a Clogston cable is to introduce a frequency dependence into the propagation constant, and in particular to cause the attenuation constant to increase, with increasing frequency, above the value which we have found for infinitesimally thin laminae (or for finite laminae at low frequencies). The increased losses are associated with the fact that the penetration depth in a laminated stack decreases with increasing frequency, even when Clogston's condition is exactly satisfied, if the laminae are of finite thickness. We shall now obtain expressions for the surface impedance of a plane laminated stack of n double layers, such as is shown in Fig. 3, when Clogston's condition is satisfied but the individual layers are of finite thickness.

We first observe that Clogston's condition (102) implies

$$\begin{aligned}\eta_{2y}\kappa_2t_2 &= \eta_2\sigma_2(1 - \gamma_0^2/\sigma_2^2)t_2 \\ &= i\omega\mu_2 \left[1 - \frac{\theta\mu_1 + (1-\theta)\mu_2}{(1-\theta)\mu_2} \right] \frac{(1-\theta)t_1}{\theta} \\ &= -i\omega\mu_1t_1 = -\eta_1\sigma_1t_1 \\ &\approx -\eta_{1y}\kappa_1t_1,\end{aligned}\tag{156}$$

where in the last step we have used the fact that in the conducting layers η_{1y} is equal to η_1 and κ_1 is equal to σ_1 to a very good approximation. We now introduce the dimensionless parameter

$$\Theta = \sigma_1t_1 = (1+i)t_1/\delta_1 \approx \kappa_1t_1,\tag{157}$$

which may be regarded as a measure of the electrical thickness of the individual conducting layers. From (86) and (156) we have, for the propagation constant per double layer,

$$\operatorname{ch} \Gamma = \operatorname{ch} \Theta - \frac{1}{2} \Theta \operatorname{sh} \Theta, \quad (158)$$

and from (87), for the iterative impedances,

$$\begin{aligned} K_1 &= \frac{\Theta}{g_1 t_1} \left[+ \frac{1}{2} \Theta + \left(\frac{1}{4} \Theta^2 - \Theta \coth \Theta + 1 \right)^{\frac{1}{2}} \right], \\ K_2 &= \frac{\Theta}{g_1 t_1} \left[- \frac{1}{2} \Theta + \left(\frac{1}{4} \Theta^2 - \Theta \coth \Theta + 1 \right)^{\frac{1}{2}} \right], \end{aligned} \quad (159)$$

since $\eta_{ly} = \kappa_1/g_1 = \Theta/g_1 t_1$.

If the thickness t_1 of each conducting layer is moderately small compared to the skin depth δ_1 at the highest frequency of interest, the quantities Γ , K_1 , and K_2 may conveniently be expanded in powers of Θ . The identity

$$\operatorname{ch} x - 1 = 2 \operatorname{sh}^2 \frac{1}{2} x \quad (160)$$

enables us to transform (158) into

$$\begin{aligned} \operatorname{sh}^2 \frac{1}{2} \Gamma &= \frac{1}{2} (\operatorname{ch} \Theta - 1) - \frac{1}{4} \Theta \operatorname{sh} \Theta \\ &= -\frac{\Theta^4}{48} \left[1 + \frac{\Theta^2}{15} + \frac{\Theta^4}{560} + \cdots \right], \end{aligned} \quad (161)$$

after we expand $\operatorname{sh} \Theta$ and $\operatorname{ch} \Theta$ by Dwight 657.1 and 657.2 and collect terms. Taking the square root by the binomial theorem gives

$$\operatorname{sh} \frac{1}{2} \Gamma = -\frac{i}{4\sqrt{3}} \left[\Theta^2 + \frac{\Theta^4}{30} + \frac{17\Theta^6}{50400} + \cdots \right], \quad (162)$$

the negative sign being introduced because from (157) Θ^2 is a positive imaginary number and we want $\operatorname{Re} \Gamma > 0$. Then

$$\begin{aligned} \Gamma &= 2 \operatorname{sh}^{-1} \left[-\frac{i}{4\sqrt{3}} \left(\Theta^2 + \frac{\Theta^4}{30} + \frac{17\Theta^6}{50400} + \cdots \right) \right] \\ &= -\frac{i}{\sqrt{3}} \left[\frac{\Theta^2}{2} + \frac{\Theta^4}{60} + \frac{\Theta^6}{525} + \cdots \right], \end{aligned} \quad (163)$$

provided that we expand the sh^{-1} function by Dwight 706. From (159) we get

$$\begin{aligned} K_1 &= \frac{1}{g_1 t_1} \left[\frac{(3 - i\sqrt{3})}{6} \Theta^2 + \frac{i\sqrt{3}}{45} \Theta^4 - \frac{i\sqrt{3}}{1575} \Theta^6 + \cdots \right], \\ K_2 &= \frac{1}{g_1 t_1} \left[-\frac{(3 + i\sqrt{3})}{6} \Theta^2 + \frac{i\sqrt{3}}{45} \Theta^4 - \frac{i\sqrt{3}}{1575} \Theta^6 + \cdots \right], \end{aligned} \quad (164)$$

where we have expanded $\coth \Theta$ by Dwight 657.5 and chosen the sign of the square root to make $\text{Re } K_1$ and $\text{Re } K_2$ both positive.

Our first observation is that when the lamina thickness is finite the effective skin depth of the stack is also finite. We have, from (157) and (163),

$$\Gamma = \frac{1}{\sqrt{3}} \left[\frac{t_1^2}{\delta_1^2} + \frac{it_1^4}{15\delta_1^4} - \frac{8t_1^6}{525\delta_1^6} - \dots \right], \quad (165)$$

and the average propagation constant per unit distance into the stack is

$$\Gamma_t = \frac{\Gamma}{(t_1 + t_2)} = \frac{1}{\sqrt{3}(t_1 + t_2)} \left[\frac{t_1^2}{\delta_1^2} + \frac{it_1^4}{15\delta_1^4} - \frac{8t_1^6}{525\delta_1^6} - \dots \right]. \quad (166)$$

If as usual we define the effective skin depth Δ to be the distance, measured into an infinitely deep stack, at which the current density has fallen to $1/e$ of its value at the surface, then keeping only the first term in (166) we have

$$\Delta = \frac{1}{\text{Re } \Gamma_t} = \frac{\sqrt{3}(t_1 + t_2)\delta_1^2}{t_1^2} = \frac{\sqrt{3}(t_1 + t_2)}{\pi\mu_1 g_1 f t_1^2}, \quad (167)$$

a result also given by Clogston.¹⁵ The number N of double layers in one effective skin depth is

$$N = \frac{\Delta}{(t_1 + t_2)} = \frac{\sqrt{3}\delta_1^2}{t_1^2} = \frac{\sqrt{3}}{\pi\mu_1 g_1 f t_1^2}, \quad (168)$$

while the total thickness T_Δ of conducting material in these layers is

$$T_\Delta = N t_1 = \frac{\sqrt{3}\delta_1^2}{t_1} = \frac{\sqrt{3}}{\pi\mu_1 g_1 f t_1}. \quad (169)$$

T_Δ is essentially the thickness of conducting material in each stack which is effectively carrying current; it is evident that for small values of t_1/δ_1 this effective thickness is inversely proportional to the frequency f and to the thickness t_1 of the individual conducting layers, but independent of the thickness t_2 of the insulating layers, provided that t_2 is very small compared to the length of a free wave in the insulating material.

In the general case, still assuming of course that Clogston's condition is satisfied, the surface impedance $Z_0(\gamma_0)$ of a plane Clogston stack is given by equation (65) of Section III, namely

$$Z_0(\gamma_0) = \frac{\frac{1}{2}Z_n(\gamma_0)(K_1 e^{n\Gamma} + K_2 e^{-n\Gamma}) + K_1 K_2 \text{sh } n\Gamma}{Z_n(\gamma_0) \text{sh } n\Gamma + \frac{1}{2}(K_1 e^{-n\Gamma} + K_2 e^{n\Gamma})}, \quad (170)$$

¹⁵ Reference 1, equation (III-44).

where $Z_n(\gamma_0)$ is the impedance of the surface behind the stack. If $\Theta = 0$, (170) reduces to (105) of Section IV, that is,

$$Z_0(\gamma_0) = \frac{1}{\bar{g}s + 1/Z_n(\gamma_0)} = \frac{1}{g_1 T_1 + 1/Z_n(\gamma_0)}, \quad (171)$$

where T_1 is the total thickness of conducting material in the stack. If $Z_n(\gamma_0)$ is infinite, then for all values of Θ and n we have

$$Z_0(\gamma_0) = \frac{\Theta}{g_1 t_1} [\tfrac{1}{2}\Theta + (\tfrac{1}{4}\Theta^2 - \Theta \coth \Theta + 1)^{\frac{1}{2}} \coth n\Gamma]; \quad (172)$$

and if $\text{Re } n\Gamma$ is large, corresponding to a stack many effective skin depths thick, then for any $Z_n(\gamma_0)$ we have

$$Z_0(\gamma_0) = K_1. \quad (173)$$

Once $Z_0(\gamma_0)$ has been computed for a particular frequency, the attenuation and phase constants of the plane Clogston 1 line at that frequency are given, as in Section II, by

$$\alpha = \text{Re } Z_0(\gamma_0)/\eta_0 b, \quad (174)$$

$$\beta = \omega \sqrt{\mu_0 \epsilon_0} + \text{Im } Z_0(\gamma_0)/\eta_0 b. \quad (175)$$

Explicit expressions for the surface impedance of a coaxial stack of finite layers have not been derived. However, if in a coaxial Clogston 1 the thickness of each stack is small compared to its mean radius, or if the depth of penetration given by (167) is small compared to the radius of the surface near which the currents flow, then the parallel-plane formula (170) may be used for the stack impedances $Z_1(\gamma_0)$ and $Z_2(\gamma_0)$ which are to be substituted into the equations of Section II for the attenuation and phase constants, namely

$$\alpha = \text{Re } \frac{Z_1(\gamma_0)/\rho_1 + Z_2(\gamma_0)/\rho_2}{2\eta_0 \log (\rho_2/\rho_1)}, \quad (176)$$

$$\beta = \omega \sqrt{\mu_0 \epsilon_0} + \text{Im } \frac{Z_1(\gamma_0)/\rho_1 + Z_2(\gamma_0)/\rho_2}{2\eta_0 \log (\rho_2/\rho_1)}. \quad (177)$$

If the plane approximations are regarded as insufficiently accurate, one can compute the surface impedance of a cylindrical stack by repeated multiplication of matrices similar to the one given by equations (88) of Section III. This procedure would obviously involve considerable numerical computation, but we can hardly expect that it would reveal anything qualitatively new for Clogston cables of the proportions considered in Part I.

It will be instructive to compare the impedance of a laminated plane stack with the impedance of a solid metal plate over the full frequency range from zero to very high frequencies.¹⁶ If the stack contains n conducting layers, each of thickness t_1 , and the metal plate is of thickness $T_1 = nt_1$, the impedances of the plate and of the stack will be equal at zero frequency, and also at very high frequencies where the first layer of the stack is already many skin depths thick. For simplicity we assume that both the plate and the stack are backed by infinite-impedance surfaces at all frequencies.

To orient ourselves we shall define three critical frequencies, for which respectively the thickness of the solid plate is equal to one skin depth in the metal, the thickness of the stack is equal to one "effective skin depth", and the thickness of a single conducting layer is equal to $\sqrt{3}$ skin depths in the metal. These frequencies are

$$\begin{aligned} f_1 &= 1/(\pi\mu_1 g_1 T_1^2) & (T_1 = \delta_1), \\ f_2 &= \sqrt{3}/(\pi\mu_1 g_1 t_1 T_1) = \sqrt{3}nf_1 & (T_1 = T_\Delta), \\ f_3 &= 3/(\pi\mu_1 g_1 t_1^2) = 3n^2 f_1 & (t_1 = \sqrt{3}\delta_1). \end{aligned} \quad (178)$$

The approximate forms of the surface impedance functions of the plate and the stack in the various frequency ranges are then quite simple.

In the range $0 \leq f \leq f_1$, the surface impedance of the solid plate is approximately constant and given by

$$Z_0(\gamma_0) \approx 1/g_1 T_1, \quad (179)$$

while in the range $f \geq f_1$ we see approximately the surface impedance of an infinite plate,

$$Z_0(\gamma_0) \approx (1 + i)/g_1 \delta_1 = (1 + i)\sqrt{\pi\mu_1 f/g_1}, \quad (180)$$

which is proportional to \sqrt{f} . The surface impedance of the stack is approximately constant in the range $0 \leq f \leq f_2$, where

$$Z_0(\gamma_0) \approx 1/g_1 T_1, \quad (181)$$

while in the range $f_2 \leq f \leq f_3$ it is approximately equal to the impedance K_1 of an infinitely deep stack of moderately thin layers as given by the first of equations (164), namely

$$Z_0(\gamma_0) \approx (1/\sqrt{3} + i)\pi\mu_1 t_1 f, \quad (182)$$

¹⁶ In this connection see also Reference 1, Fig. 2, p. 494. Clogston compares a laminated stack with a solid plate of the same total thickness as the stack, hence a plate which contains more conducting material than the stack.

which is directly proportional to frequency (and independent of conductivity). For $f \geq f_3$ the stack acts much like an infinitely thick solid plate, for which

$$Z_0(\gamma_0) \approx (1 + i)/g_1\delta_1 = (1 + i) \sqrt{\pi\mu_1 f/g_1}, \quad (183)$$

an impedance again proportional to \sqrt{f} .

The real parts of the approximate expressions for surface impedance may be plotted on log-log paper, where power-law relationships are represented by straight lines, to give quite a good idea of the way in which the stack resistance varies over the entire frequency range. To show how the exact resistance departs from the approximate formulas in the transition regions, we have calculated the resistance of a particular stack over the full frequency range from equation (172), and also the resistance of the corresponding solid plate from the formula

$$Z_0(\gamma_0) = (1 + i)\sqrt{\pi\mu_1 f/g_1} \coth [(1 + i)\sqrt{\pi\mu_1 g_1 f}T_1], \quad (184)$$

and plotted the results, together with those for an infinite plate and an infinite stack, in Fig. 7. The actual numerical values were chosen solely for ease in plotting, and are of no particular significance. It should be noted that the exact curves oscillate slightly around the asymptotic lines in the transition regions. For example, the resistance of the laminated stack is actually higher than the resistance of the solid plate at certain frequencies slightly above f_3 . These oscillations appear clearly in the numerical results, but are scarcely visible on the plots because of the logarithmic compression of the upper ends of the frequency and resistance scales.

We shall next obtain an expression for the rate at which the surface impedance of a laminated stack begins to depart from its dc value as the frequency is increased. For this purpose we must expand the various factors appearing in equation (170) for $Z_0(\gamma_0)$ in powers of Θ . Using the expansions (163) and (164) which have already been derived for Γ , K_1 , and K_2 , it is a matter of straightforward if tedious algebra to show that:

$$e^{n\Gamma} = 1 - \frac{i\sqrt{3}n}{6} \Theta^2 - \frac{(15n^2 + i2\sqrt{3}n)}{360} \Theta^4 + \dots, \quad (185)$$

$$e^{-n\Gamma} = 1 + \frac{i\sqrt{3}n}{6} \Theta^2 - \frac{(15n^2 - i2\sqrt{3}n)}{360} \Theta^4 + \dots, \quad (186)$$

$$\text{sh } n\Gamma = -\frac{in}{2\sqrt{3}} \left[\Theta^2 + \frac{\Theta^4}{30} - \frac{(175n^2 - 48)}{12600} \Theta^6 + \dots \right], \quad (187)$$

$$K_1 e^{n\Gamma} + K_2 e^{-n\Gamma} \quad (188)$$

$$= -\frac{i\Theta^2}{\sqrt{3}g_1t_1} \left[1 + \frac{(15n-4)}{30} \Theta^2 - \frac{(175n^2-70n-16)}{4200} \Theta^4 + \dots \right],$$

$$K_1 e^{-n\Gamma} + K_2 e^{n\Gamma} \quad (189)$$

$$= -\frac{i\Theta^2}{\sqrt{3}g_1t_1} \left[1 - \frac{(15n+4)}{30} \Theta^2 - \frac{(175n^2+70n-16)}{4200} \Theta^4 + \dots \right],$$

$$K_1 K_2 \operatorname{sh} n\Gamma = \frac{1}{(g_1t_1)^2} \frac{in}{6\sqrt{3}} \Theta^6 + \dots \quad (190)$$

By substituting the above series into equation (170), we can obtain the variation of the stack impedance with frequency so long as t_1/δ_1 is sufficiently small. Although in principle there would be no difficulty in taking into account an arbitrary sheath impedance $Z_n(\gamma_0)$, for brevity we shall restrict ourselves here to the case in which the sheath impedance is so high that at all frequencies of interest the current in the sheath may be neglected. Then we have equation (191) (see next page).

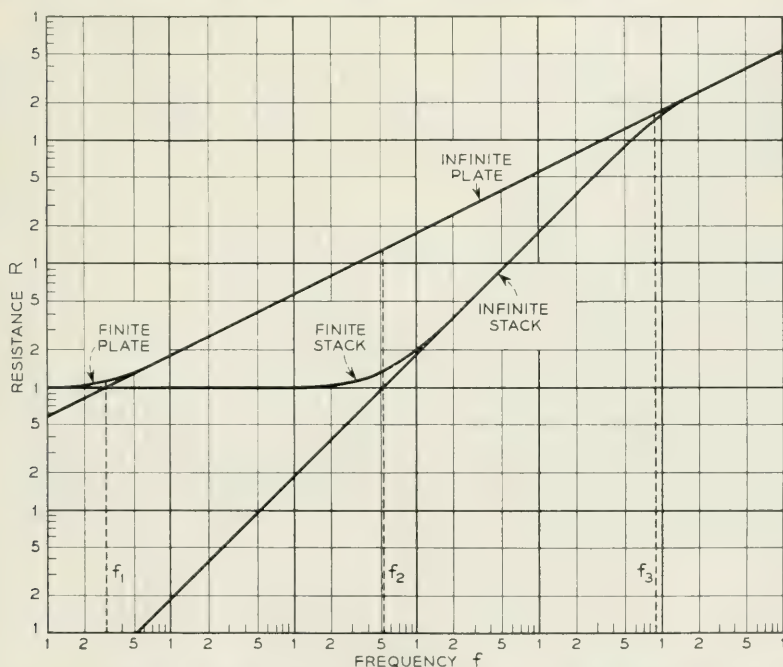


Fig. 7—Surface resistance R of solid plates and laminated stacks versus frequency f on log-log scale.

$$Z_0(\gamma_0) = \frac{K_1 e^{n\Gamma} + K_2 e^{-n\Gamma}}{2 \operatorname{sh} n\Gamma}, \quad (191)$$

which can be reduced to

$$\begin{aligned} Z_0(\gamma_0) &= \frac{1}{ng_1 t_1} \left[1 + \frac{(3n-1)}{6} \Theta^2 - \frac{(5n^2-1)}{180} \Theta^4 + \dots \right] \\ &\approx \frac{1}{g_1 T_1} \left[1 + \frac{i T_1 t_1}{\delta_1^2} + \frac{T_1^2 t_1^2}{9 \delta_1^4} + \dots \right], \end{aligned} \quad (192)$$

the last expression being valid if the number of double layers is not too small ($n \geq 5$, say). To this approximation the fractional changes in the resistance and reactance of the stack are

$$\frac{\Delta R}{R_0} = \frac{T_1^2 t_1^2}{9 \delta_1^4} = \frac{T_1^2 t_1^2 \pi^2 \mu_1^2 g_1^2 f^2}{9}, \quad (193)$$

$$\frac{\Delta X}{R_0} = \frac{T_1 t_1}{\delta_1^2} = T_1 t_1 \pi \mu_1 g_1 f, \quad (194)$$

where

$$R_0 = 1/g_1 T_1 \quad (195)$$

is the dc resistance. From the exact calculations described above it appears that (193) and (194) are valid up to the neighborhood of the critical frequency

$$f_2 = \sqrt{3}/(\pi \mu_1 g_1 t_1 T_1), \quad (196)$$

at which frequency the approximate formulas yield

$$\Delta R/R_0 = 1/3, \quad \Delta X/R_0 = \sqrt{3}. \quad (197)$$

For $f > f_2$, however, these approximations rapidly break down.

We may now answer the question: What must be the thickness t_1 of the individual conducting layers in a plane stack which contains a given total thickness T_1 of conducting material, if at a specified top frequency f_m the resistance of the stack is not to have increased by more than a specified small fraction of its dc value? We find that the permissible value of t_1 is

$$t_1 = \frac{3}{\pi \mu_1 g_1 T_1 f_m} \sqrt{\frac{\Delta R}{R_0}}, \quad (198)$$

and we note that this value of t_1 is inversely proportional both to f_m and to T_1 . If we measure t_1 and T_1 in mils and f_m in $\text{Mc} \cdot \text{sec}^{-1}$, then on putting

in the numerical values of μ_1 and g_1 for copper, we have

$$(t_1)_{\text{mils}} = \frac{20.31}{(f_m)_{\text{Mc}}(T_1)_{\text{mils}}} \sqrt{\frac{\Delta R}{R_0}}. \quad (199)$$

For a plane Clogston 1 with stacks of equal thickness, the attenuation constant is given by (174), and the fractional change in attenuation with frequency is equal to the fractional change in resistance of either stack, as calculated from (193). For a coaxial Clogston 1 with stacks thin enough so that the plane approximation is valid we may also use (193), but the fractional changes in resistance will be different for the two stacks if these are of different thicknesses, and the fractional change in the attenuation constant must be calculated from equation (176). If R_{10} and R_{20} are the dc resistances "per square" of the two stacks, and ΔR_1 and ΔR_2 their increments as obtained from (193), then the fractional increase in attenuation is given approximately by

$$\frac{\Delta\alpha}{\alpha_0} \approx \frac{\Delta R_1/\rho_1 + \Delta R_2/\rho_2}{R_{10}/\rho_1 + R_{20}/\rho_2}. \quad (200)$$

For either plane or cylindrical geometry we find that if we scale up a particular Clogston line by multiplying the thicknesses of the stacks and the main dielectric by the same factor, then the low-frequency attenuation constant will be divided by the square of the scale factor. However, the permissible thickness of the individual conducting layers, if we are to have the attenuation flat to a specified degree up to a fixed frequency, is inversely proportional to the scale factor. Thus if we double the overall dimensions of the line and double the amount of conducting material in the stacks, we shall divide the low-frequency attenuation constant by four, but we shall have to make the individual layers half as thick in order to maintain the same relative increase in attenuation constant at the same top frequency f_m . In addition it is clear that if we double the top frequency while maintaining the same requirement on $\Delta\alpha/\alpha_0$ for a line of given dimensions, we shall also have to cut the thickness of the individual layers in half.

As a numerical example, let us return to the cable whose specifications were given by (155) at the end of Section IV. For this cable we have:

$$\begin{aligned} \rho_1 &= 55.49 \text{ mils} & \theta_{s_1} &= 8.46 \text{ mils} \\ \rho_2 &= 181.44 \text{ mils} & \theta_{s_2} &= 4.04 \text{ mils} \\ \rho_2/\rho_1 &= 3.270 & R_{20}/R_{10} &\approx s_1/s_2 = 2.094 \end{aligned} \quad (201)$$

If the conducting layers are copper, we find that equation (200) for the fractional increase in attenuation becomes, numerically,

$$\Delta\alpha/\alpha_0 \approx 0.121(t_1)_{\text{mils}}^2 f_{\text{Mc}}^2. \quad (202)$$

If for example the copper layers are 0.1 mil thick and the polyethylene layers 0.05 mil thick, since we are assuming $\theta = 2/3$, then the attenuation constant has increased by 10 per cent of its "flat" value at a frequency of about $9.1 \text{ Mc} \cdot \text{sec}^{-1}$.

We may also ask for the upper crossover frequency, above which the Clogston cable will have a higher attenuation constant than a standard air-filled coaxial of the same size. Such a crossover frequency must exist because the dielectric loading of the Clogston cable (in our case $\epsilon_{0r} = 6.78$) introduces a factor $\sqrt{\epsilon_{0r}}$ into the asymptotic expression for the attenuation constant at extremely high frequencies when the stacks look like solid metal walls; in addition there will be slight differences due to the fact that the geometric proportions of the conventional and Clogston cables are not exactly the same.

We assume, subject to a posteriori verification, that the upper crossover frequency lies between the critical frequencies f_2 and f_3 , defined by (178), for each stack. Then we have in effect infinitely deep stacks of moderately thin laminae, whose surface resistances are equal and are given by (182) to be

$$R_1 = R_2 \approx \pi\mu_1 t_1 f / \sqrt{3} = 5.79 \times 10^{-5} (t_1)_{\text{mils}} f_{\text{Mc}} \text{ ohms}. \quad (203)$$

The attenuation constants of the conventional and Clogston cables are obtained from (151) and (176) respectively, where for the conventional coaxial we set $\eta_0 = \eta_v$. After a little arithmetic we find for the upper crossover frequency in this particular case,

$$f_{\text{Mc}} \approx 2.79 / (t_1)_{\text{mils}}^2. \quad (204)$$

Thus if the copper layers are 0.1 mil thick, the upper crossover frequency is about $280 \text{ Mc} \cdot \text{sec}^{-1}$, which turns out to lie well inside the interval between the critical frequencies f_2 and f_3 for both stacks.

Comparing this result with the result at the end of Section IV, we see that a 0.375-inch Clogston 1 cable with 0.1-mil copper conductors and the other specifications given by (155) is nominally better than a conventional air-filled coaxial cable of the same size in the frequency range from about $1 \text{ Mc} \cdot \text{sec}^{-1}$ to $280 \text{ Mc} \cdot \text{sec}^{-1}$. We are still neglecting the effect of failure to satisfy Clogston's condition exactly, the effect of stack non-uniformity, and dielectric losses. All of these factors will be present to a greater or less degree in any physical embodiment of a Clogston cable,

and will reduce, or in extreme cases even eliminate, the frequency range over which the Clogston cable exhibits lower loss than a conventional coaxial cable.

VI. EFFECT OF DIELECTRIC MISMATCH

We may think of Clogston's relation (102) as a condition imposed on the phase velocity in a laminated transmission line to maximize the depth of eddy current penetration into the stacks. If this condition is not exactly satisfied, that is, if the $\mu_0\epsilon_0$ product of the main dielectric is not equal to the $\bar{\mu}\bar{\epsilon}$ product of the stacks, then the effective skin depth of the stacks is finite at finite frequencies and decreases with increasing frequency even in the ideal case of infinitesimally thin layers, while if the layers are of finite thickness the effective skin depth is even less than it would be with a perfectly matched main dielectric. The losses in the stacks at moderate frequencies where Clogston's penetration effect is of importance are correspondingly increased by the presence of dielectric mismatch.

For a quantitative discussion we define the amount of dielectric mismatch $\Delta(\mu_0\epsilon_0)$ by

$$\Delta(\mu_0\epsilon_0) = \mu_0\epsilon_0 - \bar{\mu}\bar{\epsilon}, \quad (205)$$

and also the dielectric mismatch parameter k by

$$k = \frac{\Delta(\mu_0\epsilon_0)}{\bar{\mu}\bar{\epsilon} - \mu_2\epsilon_2} = \frac{(1 - \theta)}{\theta} \frac{\Delta(\mu_0\epsilon_0)}{\mu_1\epsilon_2}. \quad (206)$$

In terms of k , the general expressions for Γ , K_1 , and K_2 in a plane stack of finite layers take a relatively simple form. We have

$$\begin{aligned} \eta_{2y}\kappa_2 t_2 &= \eta_2\sigma_2(1 - \gamma_0^2/\sigma_2^2)t_2 \\ &= \frac{i\omega\mu_1}{\mu_1\epsilon_2} [\mu_2\epsilon_2 - \mu_0\epsilon_0] \frac{(1 - \theta)t_1}{\theta} \\ &= -i\omega\mu_1(1 + k)t_1 = -(1 + k)\eta_1\sigma_1 t_1 \\ &\approx -(1 + k)\eta_{1y}\kappa_1 t_1, \end{aligned} \quad (207)$$

after a little rearrangement, where the only approximation that has been made so far is to set $\eta_{1y} \approx \eta_1$ and $\kappa_1 \approx \sigma_1$. Substituting (207) into (86) and (87) gives

$$\text{ch } \Gamma = \text{ch } \Theta - \frac{1}{2}(1 + k)\Theta \text{ sh } \Theta, \quad (208)$$

and

$$K_1 = \frac{\Theta}{g_1 t_1} \left[\frac{1}{2}(1+k)\Theta + \sqrt{\frac{1}{4}(1+k)^2\Theta^2 - (1+k)\Theta \coth \Theta + 1} \right],$$

$$K_2 = \frac{\Theta}{g_1 t_1} \left[-\frac{1}{2}(1+k)\Theta + \sqrt{\frac{1}{4}(1+k)^2\Theta^2 - (1+k)\Theta \coth \Theta + 1} \right],$$
(209)

where as usual

$$\Theta = \sigma_1 t_1 = (1+i)t_1/\delta_1 \approx \kappa_1 t_1. \quad (210)$$

If $k = 0$, equations (208) and (209) evidently reduce to (158) and (159) of the preceding section. For a stack of infinitesimally thin layers, the constants Γ_ℓ and K are given by equations (93) and (94) of Section III, namely

$$\Gamma_\ell = \left[\frac{i\bar{g}}{\omega\bar{\epsilon}} (\omega^2\bar{\mu}\bar{\epsilon} - \omega^2\mu_0\epsilon_0) \right]^{\frac{1}{2}} = (-2ik)^{\frac{1}{2}}\theta/\delta_1, \quad (211)$$

$$K = \Gamma_\ell/\bar{g} = (-2ik)^{\frac{1}{2}}/g_1\delta_1. \quad (212)$$

Up to this point we have set no restrictions on the magnitude of k , and we have not even assumed that k is necessarily real. Throughout the rest of this section, however, we shall assume that k is a positive or negative real number, as it must be if there is no dielectric or magnetic dissipation.

In practice both the lamina thickness and the amount of dielectric mismatch will be as small as it is feasible to make them. It will be useful, therefore, to obtain approximate expressions for Γ , K_1 , and K_2 under the assumptions

$$|\Theta| \ll 1, \quad |k| \ll 1. \quad (213)$$

Then equation (208) yields

$$\begin{aligned} \text{sh}^2 \frac{1}{2}\Gamma &= \frac{1}{2}(\text{ch } \Theta - 1) - \frac{1}{4}(1+k)\Theta \text{sh } \Theta \\ &= -\frac{k}{4}\Theta^2 - \frac{(1+2k)}{48}\Theta^4 - \dots \end{aligned} \quad (214)$$

If $|k| \ll 1$ we can neglect $2k$ compared to unity in the coefficient of Θ^4 , but since we have made no assumptions as to the relative magnitudes of $|\Theta|$ and $|k|$, we cannot drop either the term in $k\Theta^2$ or the term in Θ^4 . If we replace $\text{sh } \frac{1}{2}\Gamma$ by $\frac{1}{2}\Gamma$ in (214), we get

$$\begin{aligned}
\Gamma &\approx [-k\Theta^2 - \Theta^4/12]^{\frac{1}{2}} \\
&= \frac{t_1}{\sqrt{3}\delta_1} [(t_1/\delta_1)^2 - 6ik]^{\frac{1}{2}} \\
&= \frac{t_1}{\sqrt{3}\delta_1} \{ [\sqrt{\frac{1}{4}(t_1/\delta_1)^4 + 9k^2} + \frac{1}{2}(t_1/\delta_1)^2]^{\frac{1}{2}} \\
&\quad - i(\operatorname{sgn} k) [\sqrt{\frac{1}{4}(t_1/\delta_1)^4 + 9k^2} - \frac{1}{2}(t_1/\delta_1)^2]^{\frac{1}{2}} \},
\end{aligned} \tag{215}$$

where we have taken the square root of the complex quantity by Dwight 58.2, and

$$\operatorname{sgn} k = \begin{cases} +1 & \text{if } k > 0, \\ -1 & \text{if } k < 0. \end{cases} \tag{216}$$

Similarly, from (209),

$$\begin{aligned}
K_1 &= \frac{1}{g_1 t_1} \left[\frac{(1+k)}{2} \Theta^2 + \Theta \sqrt{-k - \frac{(1-2k-3k^2)}{12} \Theta^2 - \dots} \right] \\
&\approx \frac{1}{g_1 t_1} \left[\frac{1}{2} \Theta^2 + \Theta \sqrt{-k - \Theta^2/12} \right] \\
&= \frac{it_1}{g_1 \delta_1^2} + \frac{1}{\sqrt{3}g_1 \delta_1} [(t_1/\delta_1)^2 - 6ik]^{\frac{1}{2}} \\
&= \frac{1}{\sqrt{3}g_1 \delta_1} \{ [\sqrt{\frac{1}{4}(t_1/\delta_1)^4 + 9k^2} + \frac{1}{2}(t_1/\delta_1)^2]^{\frac{1}{2}} + i\sqrt{3}t_1/\delta_1 \\
&\quad - i(\operatorname{sgn} k) [\sqrt{\frac{1}{4}(t_1/\delta_1)^4 + 9k^2} - \frac{1}{2}(t_1/\delta_1)^2]^{\frac{1}{2}} \}, \\
K_2 &\approx \frac{1}{\sqrt{3}g_1 \delta_1} \{ [\sqrt{\frac{1}{4}(t_1/\delta_1)^4 + 9k^2} + \frac{1}{2}(t_1/\delta_1)^2]^{\frac{1}{2}} - i\sqrt{3}t_1/\delta_1 \\
&\quad - i(\operatorname{sgn} k) [\sqrt{\frac{1}{4}(t_1/\delta_1)^4 + 9k^2} - \frac{1}{2}(t_1/\delta_1)^2]^{\frac{1}{2}} \}.
\end{aligned} \tag{217}$$

The effective skin depth of a stack of moderately thin layers in the presence of slight dielectric mismatch is, from (215),

$$\Delta = \frac{(t_1 + t_2)}{\operatorname{Re} \Gamma} = \frac{\sqrt{3}(t_1 + t_2)\delta_1/t_1}{[\sqrt{\frac{1}{4}(t_1/\delta_1)^4 + 9k^2} + \frac{1}{2}(t_1/\delta_1)^2]^{\frac{1}{2}}}. \tag{218}$$

An equation essentially equivalent to this was given by Clogston, in somewhat different notation.¹⁷ It is clear from (211) or (218) that if the layers are infinitesimally thin, we have

$$\Delta = \delta_1/\theta |k|^{\frac{1}{2}}, \tag{219}$$

and the effective skin depth in the stack is proportional to the skin

¹⁷ Reference 1, equation (III-42).

depth δ_1 in the conducting material at the operating frequency, although if the mismatch parameter k is small, the proportionality constant multiplying δ_1 will be large. In the general case, the number of double layers in one effective skin depth is

$$N = \frac{\Delta}{t_1 + t_2} = \frac{\sqrt{3}\delta_1/t_1}{[\sqrt{\frac{1}{4}(t_1/\delta_1)^4 + 9k^2 + \frac{1}{2}(t_1/\delta_1)^2}]^{\frac{1}{2}}}, \quad (220)$$

and the total thickness of conducting material in these layers is

$$T_{\Delta} = Nt_1 = \frac{\sqrt{3}\delta_1}{[\sqrt{\frac{1}{4}(t_1/\delta_1)^4 + 9k^2 + \frac{1}{2}(t_1/\delta_1)^2}]^{\frac{1}{2}}}. \quad (221)$$

It is instructive to plot the effective skin depth of a given stack at a fixed frequency as a function of dielectric mismatch. If

$$\Delta_0 = \sqrt{3}(t_1 + t_2)\delta_1^2/t_1^2 \quad (222)$$

denotes the effective skin thickness when there is no mismatch, then the relative skin thickness when the mismatch parameter is k is just

$$\frac{\Delta}{\Delta_0} = \frac{\sqrt{2}}{[\sqrt{1 + 36k^2\delta_1^4/t_1^4 + 1}]^{\frac{1}{2}}}. \quad (223)$$

This ratio is plotted against k in Fig. 8, a universal curve being obtained by measuring k in units of $(t_1/\delta_1)^2$. It is worth noting that when $k = (t_1/\delta_1)^2$, the effective skin thickness is only 53 per cent of the skin thickness with perfect dielectric match.

The surface impedance $Z_0(\gamma_0)$ of a laminated plane stack at any frequency and with any amount of dielectric mismatch is given by equation (65),

$$Z_0(\gamma_0) = \frac{\frac{1}{2}Z_n(\gamma_0)(K_1e^{n\Gamma} + K_2e^{-n\Gamma}) + K_1K_2 \operatorname{sh} n\Gamma}{Z_n(\gamma_0) \operatorname{sh} n\Gamma + \frac{1}{2}(K_1e^{-n\Gamma} + K_2e^{n\Gamma})}. \quad (224)$$

For a stack with infinitesimally thin layers and total thickness s , the equation becomes

$$Z_0(\gamma_0) = K \frac{Z_n(\gamma_0) \operatorname{ch} \Gamma_\ell s + K \operatorname{sh} \Gamma_\ell s}{Z_n(\gamma_0) \operatorname{sh} \Gamma_\ell s + K \operatorname{ch} \Gamma_\ell s}, \quad (225)$$

where Γ_ℓ and K are given by (211) and (212). At zero frequency,

$$Z_0(\gamma_0) = \frac{1}{\bar{g}s + 1/Z_n(\gamma_0)} = \frac{1}{g_1T_1 + 1/Z_n(\gamma_0)}, \quad (226)$$

while if $Z_n(\gamma_0)$ is infinite, in general

$$Z_0(\gamma_0) = \frac{\Theta}{g_1 t_1} \left\{ \frac{1}{2}(1+k)\Theta + \left[\frac{1}{4}(1+k)^2 \Theta^2 - (1+k)\Theta \coth \Theta + 1 \right]^{\frac{1}{2}} \coth n\Gamma \right\}, \quad (227)$$

which simplifies, for infinitesimally thin layers, to

$$Z_0(\gamma_0) = K \coth \Gamma t s. \quad (228)$$

If the stack is many effective skin depths thick, we have

$$Z_0(\gamma_0) = K_1, \quad (229)$$

while if the individual layers are infinitesimally thin,

$$Z_0(\gamma_0) = K, \quad (230)$$

where K_1 and K are given by (209) and (211), respectively.

When $Z_0(\gamma_0)$ is known, the attenuation and phase constants of the parallel-plane Clogston 1 are given as usual by

$$\alpha = \text{Re } Z_0(\gamma_0)/\eta_0 b, \quad (231)$$

$$\beta = \omega \sqrt{\mu_0 \epsilon_0} + \text{Im } Z_0(\gamma_0)/\eta_0 b. \quad (232)$$

For the coaxial cable we use

$$\alpha = \text{Re } \frac{Z_1(\gamma_0)/\rho_1 + Z_2(\gamma_0)/\rho_2}{2\eta_0 \log (\rho_2/\rho_1)}, \quad (233)$$

$$\beta = \omega \sqrt{\mu_0 \epsilon_0} + \text{Im } \frac{Z_1(\gamma_0)/\rho_1 + Z_2(\gamma_0)/\rho_2}{2\eta_0 \log (\rho_2/\rho_1)}, \quad (234)$$

but the impedances of the cylindrical stacks are easy to compute only if we can employ the parallel-plane approximation for each stack. To take

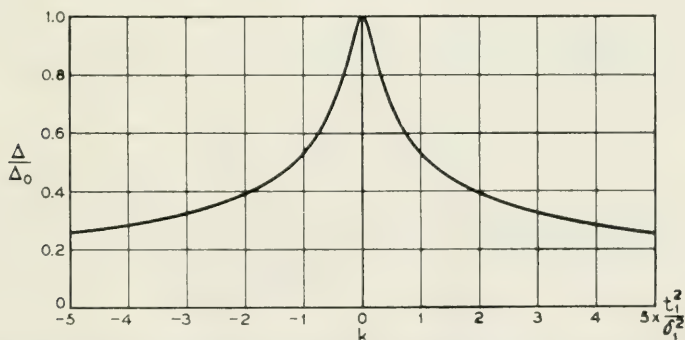


Fig. 8—Relative skin depth Δ/Δ_0 in a stack of finite layers versus dielectric mismatch parameter k , measured in units of $(t_1/\delta_1)^2$.

curvature effects into account would require a considerable amount of numerical calculation. Equation (98) of Section III provides an explicit expression for the surface impedance of a cylindrical stack of infinitesimally thin layers in the presence of dielectric mismatch, in terms of Bessel functions of complex argument; but if the layers are of finite thickness we can at present do nothing better than multiply out the matrices of the individual layers step by step.

The variation of the surface impedance of a laminated stack with frequency over the full frequency range is not quite so simple in the presence of dielectric mismatch as when Clogston's condition is exactly satisfied, but a somewhat analogous discussion may be given. As in the preceding section, we consider a plane stack of n conducting layers each of thickness t_1 , where $nt_1 = T_1$, and backed by an infinite-impedance surface. When the mismatch parameter is k , the three critical frequencies are:

$$\begin{aligned} f_1 &= 1/(\pi\mu_1 g_1 T_1^2) & (T_1 = \delta_1), \\ f_2 &= \sqrt{3}/(\pi\mu_1 g_1 t_1 T_1 \sqrt{1 + 3n^2 k^2}) \\ &= \sqrt{3} n f_1 / \sqrt{1 + 3n^2 k^2} & (T_1 = T_\Delta), \\ f_3 &= 3/(\pi\mu_1 g_1 t_1^2) = 3n^2 f_1 & (t_1 = \sqrt{3}\delta_1). \end{aligned} \quad (235)$$

In the range $0 \leq f \leq f_2$, the surface impedance of the stack is approximately constant, being given by

$$Z_0(\gamma_0) \approx 1/g_1 T_1. \quad (236)$$

In the range $f_2 \leq f \leq f_3$, we have

$$Z_0(\gamma_0) \approx K_1, \quad (237)$$

where K_1 is given by (217) provided that k is small compared to unity. For infinitesimally thin layers the upper critical frequency f_3 is infinite, and we have for $f \geq f_2$,

$$\begin{aligned} Z_0(\gamma_0) &\approx |k|^{\frac{1}{2}}(1 - i \operatorname{sgn} k)/g_1 \delta_1 \\ &= (1 - i \operatorname{sgn} k) \sqrt{\pi\mu_1 |k| f/g_1}, \end{aligned} \quad (238)$$

which is proportional to \sqrt{f} . If the layers are of finite thickness but $k = 0$, we have the result obtained in the preceding section,

$$Z_0(\gamma_0) \approx (1/\sqrt{3} + i)\pi\mu_1 t_1 f, \quad (239)$$

which is proportional to f up to the critical frequency f_2 . If neither the mismatch parameter k nor the layer thickness t_1 is zero, then the surface

impedance $Z_0(\gamma_0)$ cannot be represented by a simple power of f in the range $f_2 \leq f \leq f_3$. At frequencies above f_3 , if the layer thickness is finite, the impedance is approximately that of a solid conductor, namely

$$Z_0(\gamma_0) \approx (1 + i)/g_1\delta_1 = (1 + i)\sqrt{\pi\mu_1 f/g_1}, \quad (240)$$

which is proportional to \sqrt{f} .

Since in general the surface resistance depends upon the two parameters t_1/δ_1 and k , it is not possible to plot a single curve which shows the variation of resistance with frequency under all possible conditions of dielectric mismatch. However if we compare a matched stack of finite layers with a similar mismatched stack, we see that the asymptotic behavior of $Z_0(\gamma_0)$ is the same for both stacks at very low and very high frequencies. A numerical study of the exact equation for $Z_0(\gamma_0)$ shows that in the neighborhood of the critical frequency f_2 , the resistance of the mismatched stack is higher than the resistance of the matched stack. (The critical frequency f_2 as defined in (235) is a function of the mismatch parameter k , but will be of the same order of magnitude for a slightly mismatched stack as for a perfectly matched stack.) The resistance of the mismatched stack exhibits relatively small fluctuations above and below the resistance of the matched stack in the neighborhood of the upper critical frequency f_3 , but this region is not of as much practical interest as the region near f_2 , where the stack resistance is definitely increased by the effect of dielectric mismatch.

An explicit expression for the rate at which the surface impedance of a mismatched stack begins to depart from its dc value as the frequency is increased has been worked out only for the ideal case of infinitesimally thin layers. For a plane stack of infinitesimal layers backed by an infinite-impedance surface, equation (228) gives, at moderately low frequencies,

$$\begin{aligned} Z_0(\gamma_0) &= \frac{K}{\Gamma_{ts}} \left[1 + \frac{(\Gamma_{ts})^2}{3} - \frac{(\Gamma_{ts})^4}{45} + \dots \right] \\ &= \frac{1}{g_1 T_1} \left[1 - \frac{2ikT_1^2}{3\delta_1^2} + \frac{4k^2 T_1^4}{45\delta_1^4} + \dots \right], \end{aligned} \quad (241)$$

from which the fractional changes in resistance and reactance are

$$\frac{\Delta R}{R_0} = \frac{4k^2 T_1^4}{45\delta_1^4} = \frac{4k^2 \pi^2 \mu_1^2 g_1^2 T_1^4 f^2}{45}, \quad (242)$$

$$\frac{\Delta X}{R_0} = -\frac{2kT_1^2}{3\delta_1^2} = -\frac{2k\pi\mu_1 g_1 T_1^2 f}{3}. \quad (243)$$

The admissible value of $|k|$, if the fractional change in resistance is not to exceed a specified value $\Delta R/R_0$ at a given top frequency f_m , is

$$|k| = \frac{3\sqrt{5}\delta_1^2}{2T_1^2} \sqrt{\frac{\Delta R}{R_0}} = \frac{3\sqrt{5}}{2\pi\mu_1 g_1 f_m T_1^2} \sqrt{\frac{\Delta R}{R_0}}, \quad (244)$$

which is inversely proportional both to f_m and to the square of the total thickness of conducting material in the stack. If we express T_1 in mils, f_m in $\text{Mc} \cdot \text{sec}^{-1}$, and assume the conducting layers to be copper, we get

$$|k| = \frac{22.71}{(f_m)_{\text{Mc}} (T_1)_{\text{mils}}^2} \sqrt{\frac{\Delta R}{R_0}}. \quad (245)$$

The variation with frequency of the surface impedance of a matched stack of finite layers at moderate frequencies (say $f \leq f_2$) is given by equation (192) of Section V; but no simple formula has yet been derived for the surface impedance of a mismatched stack of finite layers in this frequency range. The derivation of such a formula would appear to involve nothing more than some rather formidable algebra, the difficulties centering around the fact that in the general case we can make no a priori assumptions as to the relative magnitudes of k and $(t_1/\delta_1)^2$. It is reasonable to suppose, however, that if both dielectric mismatch and finite lamina thickness contribute appreciably to $\Delta R/R_0$, the permissible values of $|k|$ and t_1 individually will be less, if we are to achieve a given flatness of the attenuation versus frequency curve, than the permissible value of either if the other factor were unimportant.

To exhibit the effect of dielectric mismatch from a slightly different point of view, we may plot the surface resistance of an infinitely deep plane stack of moderately thin layers (a finite stack several effective skin depths thick would show essentially the same behavior) at a fixed frequency, as a function of the mismatch parameter k . The surface resistance is just $\text{Re } K_1$, which may be obtained from (217) if k and t_1/δ_1 are assumed small compared to unity. Fig. 9 shows the dimensionless quantity

$$\text{Re } g_1 \delta_1 K_1 = \frac{1}{\sqrt{3}} \left[\sqrt{\frac{1}{4}(t_1/\delta_1)^4 + 9k^2} + \frac{1}{2}(t_1/\delta_1)^2 \right]^{\frac{1}{2}}, \quad (246)$$

for the three values $t_1/\delta_1 = 0$, $t_1/\delta_1 = 0.1$, and $t_1/\delta_1 = 0.2$. For an electrically thick solid conductor we have simply

$$\text{Re } g_1 \delta_1 K_1 = 1; \quad (247)$$

hence to get any benefit from the laminated stack we must have $\text{Re } g_1 \delta_1 K_1$ smaller than unity. Actually, if we meet Clogston's condition by

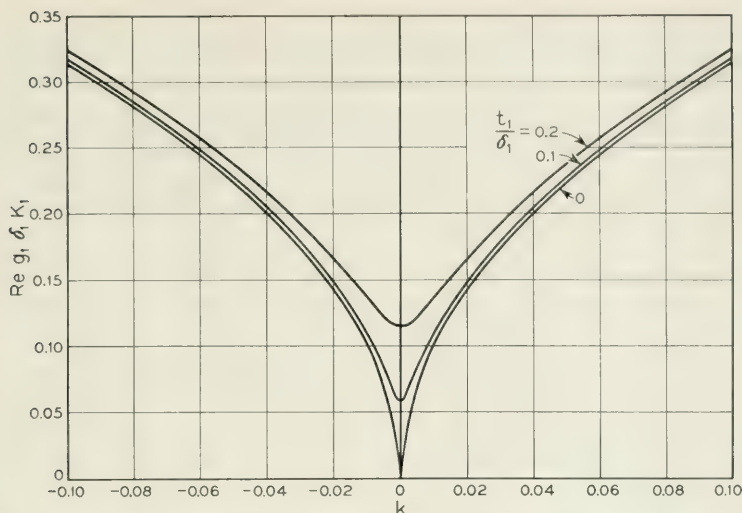


Fig. 9—Normalized stack resistance $\text{Re } g_1 \delta_1 K_1$ versus dielectric mismatch parameter k , for different values of t_1/δ_1 .

raising the dielectric constant and thus lowering the impedance of the main dielectric, then since the attenuation constant of the line is proportional to the ratio of stack resistance to dielectric impedance, we must have $\text{Re } g_1 \delta_1 K_1$ considerably smaller than unity to obtain a lower attenuation with the Clogston line than with an ordinary air-filled line having solid metal walls.

For a plane Clogston 1 line with stacks of equal thickness, the fractional change in the attenuation constant with frequency is equal to the fractional change in the resistance of either stack, whether this change arises from the effects of finite lamina thickness or from dielectric mismatch or both. The fractional change in the attenuation constant of a coaxial Clogston 1 depends not only on the change in resistance of each stack, but also on the geometric proportions of the cable, in the manner expressed by equation (200) of Section V.

The effect of dielectric mismatch on the overall attenuation versus frequency characteristic of a Clogston cable is in general to reduce the total frequency range (in Mc-sec^{-1}) over which the Clogston cable has a smaller attenuation constant than a conventional air-filled coaxial cable of the same size. To calculate the lower crossover frequency we may ordinarily neglect finite lamina thickness effects and use equation (241) for the stack impedances, while at the upper crossover frequency the stack impedances are very nearly equal to K_1 , as given by (217).

It should be remembered that mismatch of the $\mu_0\epsilon_0$ product of the main dielectric will usually be accompanied by a change in the dielectric impedance $\sqrt{\mu_0/\epsilon_0}$. Thus under certain conditions the lower crossover frequency may even be reduced by choosing ϵ_0 slightly below the Clogston value, inasmuch as the increase in dielectric impedance may more than compensate for the increase in stack resistance at low frequencies; but it appears that this will be paid for in a steeper slope of the attenuation versus frequency curve and a consequent greater reduction of the upper crossover frequency.

It would be very useful to make a numerical study of the effects of dielectric mismatch in Clogston cables having a variety of different proportions; but in the present paper space limitations restrict us to a few observations concerning orders of magnitude. For the cable which we considered at the end of the preceding section, it turns out that an increase or decrease of 1 per cent in the value of ϵ_0 makes a change of at most a very few per cent in either crossover frequency; with a matched dielectric, we recall, these crossover frequencies were about $1 \text{ Mc} \cdot \text{sec}^{-1}$ and about $280 \text{ Mc} \cdot \text{sec}^{-1}$ respectively. However if we had designed a laminated cable with thicker stacks or thinner laminae or both, so as to increase the theoretical factor of improvement over a conventional cable in the working frequency range, we should have found that the tolerable deviation of ϵ_0 from Clogston's value, instead of being of the order of 1 per cent, was more nearly of the order of 0.1 per cent or even smaller; and the greater the improvement striven for, the more stringent the requirement of accurate dielectric match.

VII. DIELECTRIC AND MAGNETIC LOSSES IN CLOGSTON 1 LINES

Dielectric and magnetic dissipation in the main dielectric and in the stacks can be taken into account by introducing complex dielectric constants and permeabilities for the lossy materials. Thus we may write

$$\begin{aligned}\epsilon_0 &= \epsilon'_0 - i\epsilon''_0 = \epsilon'_0 (1 - i \tan \phi_0), \\ \epsilon_2 &= \epsilon'_2 - i\epsilon''_2 = \epsilon'_2 (1 - i \tan \phi_2), \\ \mu_0 &= \mu'_0 - i\mu''_0 = \mu'_0 (1 - i \tan \zeta_0), \\ \mu_1 &= \mu'_1 - i\mu''_1 = \mu'_1 (1 - i \tan \zeta_1), \\ \mu_2 &= \mu'_2 - i\mu''_2 = \mu'_2 (1 - i \tan \zeta_2),\end{aligned}\tag{248}$$

where in the most general case the loss tangents may all be different, though it will be assumed that they are all small compared to unity, so that the problem may be treated by first-order perturbation methods.

The average rate of energy dissipation per unit volume in a lossy dielectric by a harmonically varying electric field of maximum amplitude E is just $\frac{1}{2}\omega\epsilon''E^2$, since the imaginary part ϵ'' of the complex dielectric constant corresponds to a conductivity $g = \omega\epsilon''$. Similarly the average rate of energy dissipation per unit volume in a lossy magnetic material by a harmonically varying magnetic field of maximum amplitude H is $\frac{1}{2}\omega\mu''H^2$. The part of the attenuation constant which arises from dielectric and magnetic dissipation is one-half the ratio of power dissipated per unit length of line to total transmitted power, provided of course that the attenuation per wavelength is small. Since the loss tangents of the various materials are assumed small, we can use the fields found for the lossless case to calculate the transmitted and dissipated power.

If the volume occupied by currents in the stacks is small compared to the volume of the main dielectric, so that we can neglect the power flow in the stacks in the direction of wave propagation compared to the power flow in the main dielectric, then the part of the attenuation constant which is due to dielectric and magnetic dissipation is given by equation (51) of Section II, namely

$$\alpha_d = \frac{1}{2}\omega\sqrt{\mu'_0\epsilon'_0}(\tan\phi_0 + \tan\zeta_0) = \frac{\pi\sqrt{\mu'_{0r}\epsilon'_{0r}}}{\lambda_v}(\tan\phi_0 + \tan\zeta_0), \quad (249)$$

where λ_v is the vacuum wavelength and μ'_{0r} , ϵ'_{0r} are the real parts of the relative permeability and relative dielectric constant of the main dielectric. This equation will be derived from energy considerations presently. It should be noted that the part of the attenuation constant given by (249) is directly proportional to frequency, provided that the loss tangents are independent of frequency; but it is the same for both plane and coaxial geometry and is independent of all the geometrical factors which describe the size and the relative proportions of the line.

Equation (249) will probably be sufficiently accurate for all Clogston 1 lines having the proportions (stacks thin compared to main dielectric) which we have considered in Part I. As an example wherein we also take into account the power flow in the stacks, however, we shall treat a parallel-plane line with infinitesimally thin laminae backed by high-impedance walls. Then, according to equations (120) and (121) of Section IV, the principal field components in the main dielectric are

$$\begin{aligned} H_x &\approx H_0, \\ E_y &\approx -\sqrt{\mu'_0/\epsilon'_0}H_0, \end{aligned} \quad (250)$$

and in the stacks,

$$\begin{aligned} H_x &\approx H_0(\tfrac{1}{2}a \mp y)/s, \\ \bar{E}_y &\approx -\sqrt{\bar{\mu}'/\bar{\epsilon}'} H_0(\tfrac{1}{2}a \mp y)/s, \end{aligned} \quad (251)$$

the propagation factor $e^{-\gamma z + i\omega t}$ being understood throughout. To take account of dielectric and magnetic dissipation in the stacks, we write

$$\begin{aligned} \bar{\epsilon} &= \bar{\epsilon}' - i\bar{\epsilon}'' = [\epsilon'_2/(1-\theta)] - i[\epsilon''_2/(1-\theta)], \\ \bar{\mu} &= \bar{\mu}' - i\bar{\mu}'' = [\theta\mu'_1 + (1-\theta)\mu'_2] - i[\theta\mu''_1 + (1-\theta)\mu''_2]. \end{aligned} \quad (252)$$

The average power P_0 transmitted through the main dielectric is obtained by integrating the real part of the z -component of the complex Poynting vector $\frac{1}{2}\mathbf{E} \times \mathbf{H}^*$ over unit width of the line; thus

$$P_0 = \frac{1}{2} \int_{-\frac{1}{2}b}^{\frac{1}{2}b} \sqrt{\mu'_0/\epsilon'_0} H_0 H_0^* dy = \frac{1}{2} \sqrt{\mu'_0/\epsilon'_0} H_0 H_0^* b. \quad (253)$$

Similarly, the average power P_1 transmitted per unit width of either stack is

$$\begin{aligned} P_1 &= \frac{1}{2} \int_{\frac{1}{2}b}^{\frac{1}{2}a} [\sqrt{\bar{\mu}'/\bar{\epsilon}'} H_0 H_0^* (\tfrac{1}{2}a - y)^2/s^2] dy \\ &= \frac{1}{6} \sqrt{\bar{\mu}'/\bar{\epsilon}'} H_0 H_0^* s. \end{aligned} \quad (254)$$

The average power ΔP_0 dissipated in the main dielectric per unit length and width of the line is

$$\begin{aligned} \Delta P_0 &= \frac{1}{2}\omega \int_{-\frac{1}{2}b}^{\frac{1}{2}b} [\epsilon''_0 E_y E_y^* + \mu''_0 H_x H_x^*] dy \\ &= \frac{1}{2}\omega [\epsilon''_0 (\mu'_0/\epsilon'_0) + \mu''_0] H_0 H_0^* b \\ &= \frac{1}{2}\omega \mu'_0 H_0 H_0^* b (\tan \phi_0 + \tan \zeta_0), \end{aligned} \quad (255)$$

while the average power ΔP_1 dissipated per unit length and width of either stack is

$$\begin{aligned} \Delta P_1 &= \frac{1}{2}\omega \int_{\frac{1}{2}b}^{\frac{1}{2}a} [\bar{\epsilon}'' \bar{E}_y \bar{E}_y^* + \bar{\mu}'' H_x H_x^*] dy \\ &= \frac{1}{6}\omega \bar{\mu}' H_0 H_0^* s (\tan \phi_2 + \tan \bar{\zeta}), \end{aligned} \quad (256)$$

where

$$\tan \bar{\zeta} = \frac{\bar{\mu}''}{\bar{\mu}'} = \frac{\theta\mu''_1 + (1-\theta)\mu''_2}{\theta\mu'_1 + (1-\theta)\mu'_2}. \quad (257)$$

The attenuation constant due to dielectric and magnetic dissipation is

$$\alpha_d = \frac{\Delta P_0 + 2\Delta P_1}{2(P_0 + 2P_1)} \quad (258)$$

$$= \frac{1}{2}\omega\sqrt{\mu'_0\epsilon'_0} \frac{(\tan\phi_0 + \tan\xi_0) + (2\bar{\mu}'s/3\mu'_0b)(\tan\phi_2 + \tan\bar{\xi})}{1 + (2s/3b)\sqrt{\bar{\mu}'\epsilon'_0/\mu'_0\bar{\epsilon}'}} ,$$

which reduces to (249) if we neglect the terms in s/b . The total attenuation is the sum of the metal losses, given by equation (110), and the dielectric and magnetic losses.

For a coaxial Clogston cable with infinitesimally thin laminae and high-impedance boundaries, the principal field components are given by equations (126)–(128) of Section IV. In the main dielectric we have

$$H_\phi \approx \frac{I}{2\pi\rho}, \quad (259)$$

$$E_\rho \approx \sqrt{\frac{\mu'_0}{\epsilon'_0}} \frac{I}{2\pi\rho},$$

while in the inner stack,

$$H_\phi \approx \frac{I(\rho^2 - a^2)}{2\pi\rho(\rho_1^2 - a^2)}, \quad (260)$$

$$\bar{E}_\rho \approx \sqrt{\frac{\bar{\mu}'}{\bar{\epsilon}'}} \frac{I(\rho^2 - a^2)}{2\pi\rho(\rho_1^2 - a^2)},$$

and in the outer stack,

$$H_\phi \approx \frac{I(b^2 - \rho^2)}{2\pi\rho(b^2 - \rho_2^2)}, \quad (261)$$

$$\bar{E}_\rho \approx \sqrt{\frac{\bar{\mu}'}{\bar{\epsilon}'}} \frac{I(b^2 - \rho^2)}{2\pi\rho(b^2 - \rho_2^2)}.$$

The average power transmitted through the main dielectric is

$$P_0 = \frac{1}{2} \sqrt{\frac{\mu'_0}{\epsilon'_0}} \frac{II^*}{2\pi} \log \frac{\rho_2}{\rho_1}, \quad (262)$$

while for the average power transmitted through the inner and outer stacks it will be sufficient to replace the exact expressions by the following simple approximations,

$$P_1 \approx \frac{1}{2} \sqrt{\frac{\bar{\mu}'}{\bar{\epsilon}'}} \frac{II^*}{2\pi} \frac{s_1}{3\rho_1}, \quad (263)$$

$$P_2 \approx \frac{1}{2} \sqrt{\frac{\bar{\mu}'}{\bar{\epsilon}'}} \frac{II^*}{2\pi} \frac{s_2}{3\rho_2}. \quad (264)$$

For the average power dissipated per unit length of line in the main dielectric and the inner and outer stacks we have, respectively,

$$\Delta P_0 = \frac{1}{2} \omega \mu_0' \frac{II^*}{2\pi} \log \frac{\rho_2}{\rho_1} (\tan \phi_0 + \tan \zeta_0), \quad (265)$$

$$\Delta P_1 \approx \frac{1}{2} \omega \bar{\mu}' \frac{II^*}{2\pi} \frac{s_1}{3\rho_1} (\tan \phi_2 + \tan \bar{\zeta}), \quad (266)$$

$$\Delta P_2 \approx \frac{1}{2} \omega \bar{\mu}' \frac{II^*}{2\pi} \frac{s_2}{3\rho_2} (\tan \phi_2 + \tan \bar{\zeta}).$$

The part of the attenuation constant which is due to dielectric and magnetic dissipation is therefore

$$\begin{aligned} \alpha_d &= \frac{\Delta P_0 + \Delta P_1 + \Delta P_2}{2(P_0 + P_1 + P_2)} \\ &= \frac{\frac{1}{2} \omega \sqrt{\mu_0' \epsilon_0'} \log \frac{\rho_2}{\rho_1} (\tan \phi_0 + \tan \zeta_0) + \frac{1}{3} \frac{\bar{\mu}'}{\mu_0'} \left(\frac{s_1}{\rho_1} + \frac{s_2}{\rho_2} \right) (\tan \phi_2 + \tan \bar{\zeta})}{\log \frac{\rho_2}{\rho_1} + \frac{1}{3} \sqrt{\frac{\bar{\mu}' \epsilon_0'}{\mu_0' \epsilon_0'}} \left(\frac{s_1}{\rho_1} + \frac{s_2}{\rho_2} \right)}. \end{aligned}$$

We need scarcely point out that if the loss tangents are not small compared to unity, it may be impossible to satisfy Clogston's condition (102) very closely with a real value of θ , and the resulting mismatch may reduce the depth of penetration and increase the metal losses in the stacks. In practice, however, the loss tangents will be of the order of 0.001 or even 0.0001, and matching the imaginary parts of $\mu_0 \epsilon_0$ and $\bar{\mu} \bar{\epsilon}$ will be much less of a practical problem than matching the real parts.

APPENDIX I

BESSEL FUNCTION EXPANSIONS

Let ρ_1 and ρ_2 be the inner and outer radii of a cylindrical shell and let the thickness t , given by

$$t = \rho_2 - \rho_1, \quad (A1)$$

be less than ρ_1 . Then, following Schelkunoff,¹ we may replace the Bessel functions appearing in equation (68) of Section III by their Taylor expansions, namely

¹ S. A. Schelkunoff, *Bell System Tech. J.*, **13**, pp. 561-562 (1934).

$$\begin{aligned}
 I_0(\kappa\rho_2) &= I_0(\kappa\rho_1 + \kappa l) = \sum_{n=0}^{\infty} \frac{(\kappa l)^n}{n!} I_0^{(n)}(\kappa\rho_1), \\
 K_0(\kappa\rho_2) &= K_0(\kappa\rho_1 + \kappa l) = \sum_{n=0}^{\infty} \frac{(\kappa l)^n}{n!} K_0^{(n)}(\kappa\rho_1), \\
 I_1(\kappa\rho_2) &= I_0'(\kappa\rho_2) = \sum_{n=0}^{\infty} \frac{(\kappa l)^n}{n!} I_0^{(n+1)}(\kappa\rho_1), \\
 K_1(\kappa\rho_2) &= -K_0'(\kappa\rho_2) = -\sum_{n=0}^{\infty} \frac{(\kappa l)^n}{n!} K_0^{(n+1)}(\kappa\rho_1).
 \end{aligned} \tag{A2}$$

It follows that

$$\begin{aligned}
 K_0(\kappa\rho_1)I_1(\kappa\rho_2) + K_1(\kappa\rho_2)I_0(\kappa\rho_1) &= -\sum_{n=0}^{\infty} \frac{(\kappa l)^n}{n!} B_{n+1}(\kappa\rho_1), \\
 K_0(\kappa\rho_1)I_0(\kappa\rho_2) - K_0(\kappa\rho_2)I_0(\kappa\rho_1) &= -\sum_{n=0}^{\infty} \frac{(\kappa l)^n}{n!} B_n(\kappa\rho_1), \\
 K_1(\kappa\rho_1)I_1(\kappa\rho_2) - K_1(\kappa\rho_2)I_1(\kappa\rho_1) &= \sum_{n=0}^{\infty} \frac{(\kappa l)^n}{n!} A_{n+1}(\kappa\rho_1), \\
 K_1(\kappa\rho_1)I_0(\kappa\rho_2) + K_0(\kappa\rho_2)I_1(\kappa\rho_1) &= \sum_{n=0}^{\infty} \frac{(\kappa l)^n}{n!} A_n(\kappa\rho_1),
 \end{aligned} \tag{A3}$$

where

$$\begin{aligned}
 A_n(x) &= I_0'(x)K_0^{(n)}(x) - K_0'(x)I_0^{(n)}(x), \\
 B_n(x) &= I_0(x)K_0^{(n)}(x) - K_0(x)I_0^{(n)}(x).
 \end{aligned} \tag{A4}$$

The quantities $A_n(x)$ and $B_n(x)$ turn out to be finite polynomials in $1/x$, the general expressions for the coefficients having been derived in a rather inaccessible monograph by Pleijel.² When x is large, however, the leading terms are quite simple. From Pleijel's analysis, or directly by substituting the asymptotic series for $I_0(x)$ and $K_0(x)$ into (A4), we find

$$\begin{aligned}
 A_{2m}(x) &= 1/x + O(1/x^3), \\
 A_{2m+1}(x) &= -m/x^2 + O(1/x^4), \\
 B_{2m}(x) &= m/x^2 + O(1/x^4), \\
 B_{2m+1}(x) &= -1/x + O(1/x^3),
 \end{aligned} \tag{A5}$$

where m is a positive integer or zero.

If we substitute these approximations into the first of equations (A3), we obtain

² H. Pleijel, *Beräkning af Motstånd och Själfinduktion*, K. L. Beckmans Boktryckeri, Stockholm, 1906.

$$\begin{aligned}
& K_0(\kappa\rho_1)I_1(\kappa\rho_2) + K_1(\kappa\rho_2)I_0(\kappa\rho_1) \\
& \approx \frac{1}{\kappa\rho_1} \sum_{m=0}^{\infty} \frac{(\kappa t)^{2m}}{(2m)!} - \frac{1}{(\kappa\rho_1)^2} \sum_{m=0}^{\infty} \frac{(m+1)(\kappa t)^{2m+1}}{(2m+1)!} \\
& = \frac{1}{\kappa\rho_2} \operatorname{ch} \kappa t - \frac{1}{(\kappa\rho_1)^2} \left[\frac{1}{2} \frac{d}{dx} (x \operatorname{sh} x) \right]_{x=\kappa t} \\
& = \left[\frac{1}{\kappa\rho_1} - \frac{t}{2\kappa\rho_1^2} \right] \operatorname{ch} \kappa t - \frac{1}{2(\kappa\rho_1)^2} \operatorname{sh} \kappa t.
\end{aligned} \tag{A6}$$

The other three equations may be treated similarly. Doing so, and remembering that

$$\rho_2/\rho_1 = 1 + t/\rho_1, \tag{A7}$$

we obtain the results which were quoted in Section III, namely

$$\begin{aligned}
\kappa\rho_2(K_{01}I_{12} + K_{12}I_{01}) & \approx \left[1 + \frac{t}{2\rho_1} \right] \operatorname{ch} \kappa t - \frac{1}{2\kappa\rho_1} \operatorname{sh} \kappa t, \\
\kappa\rho_2(K_{01}I_{02} - K_{02}I_{01}) & \approx \left[1 + \frac{t}{2\rho_1} \right] \operatorname{sh} \kappa t, \\
\kappa\rho_2(K_{11}I_{12} - K_{12}I_{11}) & \approx \left[1 + \frac{t}{2\rho_1} \right] \operatorname{sh} \kappa t, \\
\kappa\rho_2(K_{11}I_{02} + K_{02}I_{11}) & \approx \left[1 + \frac{t}{2\rho_1} \right] \operatorname{ch} \kappa t + \frac{1}{2\kappa\rho_1} \operatorname{sh} \kappa t,
\end{aligned} \tag{A8}$$

up to first order in t/ρ_1 .

TABLE OF SYMBOLS

Note: Rationalized MKS units are employed throughout. The subscripts 0, 1, 2 applied to symbols representing material constants, such as ϵ , μ , g , σ , and η , have the significance that 0 refers to the main dielectric in a Clogston line, while 1 refers to the conducting layers and 2 refers to the insulating layers in the stacks. Bars denote average values. Subscripts not included in the present table are explained in the context where they are used.

α , β , ϵ , \mathfrak{D} : Elements of the general circuit parameter matrix (Section III).

a : Distance between outer sheaths of plane Clogston line.
Radius of inner core of coaxial Clogston line.

b : Thickness of main dielectric in plane Clogston line. Inner radius of outer sheath of coaxial Clogston line.

C :	A parameter related to the degree of nonuniformity in a laminated medium (Section XII).
E :	Electric field intensity; coordinate subscripts indicate components.
f :	Frequency.
g :	Electrical conductivity.
\tilde{g} :	θg_1 ; average conductivity parallel to laminated stack.
H :	Magnetic field intensity; coordinate subscripts indicate components.
h :	$-ik_0$; a transverse separation constant (Section X).
I :	Electric current.
i :	$\sqrt{-1}$.
J :	Electric current density; coordinate subscripts indicate components.
K :	Characteristic impedance of stack of infinitesimally thin laminae.
K_1, K_2 :	Characteristic or iterative impedances of laminated stack (introduced in Section III).
k :	A parameter related to dielectric mismatch in a Clogston 1 line (Section VI).
M :	The general circuit parameter matrix ($\mathcal{A}\mathcal{B}\mathcal{C}\mathcal{D}$ -matrix).
m :	A mode number.
n :	Number of double layers in a laminated stack.
p :	A mode number.
q :	A parameter related to the propagation constant in a Clogston 2 line (Section XI).
R :	A-c resistance of a laminated stack.
r :	Ratio of attenuation constants of Clogston and conventional lines (Section XII).
s :	Thickness of a laminated stack.
s_1, s_2 :	Thicknesses of inner and outer stacks in a coaxial Clogston 1.
T :	Total thickness of conducting material in a laminated stack (subscripts explained in context).
T_Δ :	Total thickness of conducting material in one effective skin depth.
t :	Thickness of an electrically homogeneous layer. Time.
t_1 :	Thickness of a single conducting layer.
t_2 :	Thickness of a single insulating layer.
V :	Electric potential.
w :	An abbreviation for H_y in Section XII.

X :	AC reactance of a laminated stack.
x :	Rectangular coordinate in the direction of magnetic field in a plane Clogston line.
y :	Rectangular coordinate in the direction normal to the stacks in a plane Clogston line.
Z :	Surface impedance of a plane or cylindrical boundary; ratio of tangential components of the electric and magnetic fields (subscripts explained in context).
Z_k :	Characteristic impedance of a transmission line.
z :	Rectangular coordinate in the direction of wave propagation.
α :	Re γ ; attenuation constant.
β :	Im γ ; phase constant.
Γ :	Propagation constant per double layer normal to laminated stack.
Γ_t :	$\Gamma/(t_1 + t_2)$; average propagation constant per unit distance normal to laminated stack.
γ :	Propagation constant in longitudinal direction.
Δ :	Effective skin depth; the depth at which the current density in an infinite plane stack has fallen to $1/e$ of its value at the surface. A small change in a quantity.
δ :	$\sqrt{2/\omega\mu g}$; skin thickness in a solid conductor.
ϵ :	Dielectric constant (capacitivity or permittivity).
$\bar{\epsilon}$:	$\epsilon_2/(1 - \theta)$; average dielectric constant measured normal to laminated stack.
ϵ_r :	ϵ/ϵ_v ; relative dielectric constant.
ϵ_v :	Dielectric constant of vacuum; 8.854×10^{-12} farads·meter ⁻¹ .
ϵ', ϵ'' :	Real and (negative) imaginary parts of complex dielectric constant.
ζ :	$\tan^{-1}(\mu''/\mu')$; phase angle of complex permeability.
η :	$\sqrt{i\omega\mu/(g + i\omega\epsilon)}$; intrinsic impedance of medium.
η_v :	Intrinsic impedance of vacuum; 376.7 ohms.
η_y, η_ρ :	$\eta(1 - \gamma^2/\sigma^2)^{\frac{1}{2}}$; characteristic impedance looking in the y - or ρ -direction in a homogeneous medium.
Θ :	$(1 + i)t_1/\delta_1$; a parameter related to the electrical thickness of a conducting layer.
θ :	$t_1/(t_1 + t_2)$; fraction of stack volume filled by conducting layers.
κ :	$(\sigma^2 - \gamma^2)^{\frac{1}{2}}$; transverse propagation constant in the y - or ρ -direction in a homogeneous medium.

Δ :	A parameter related to the propagation constant in a Clogston 2 (Section XII).
λ :	Wavelength.
λ_v :	Free-space wavelength.
μ :	Permeability.
$\bar{\mu}$:	$\theta\mu_1 + (1 - \theta)\mu_2$; average permeability measured parallel to laminated stack.
μ_r :	μ/μ_v ; relative permeability.
μ_v :	Permeability of vacuum; $4\pi \times 10^{-7}$ henrys·meter ⁻¹ .
μ', μ'' :	Real and (negative) imaginary parts of complex permeability.
ξ :	$y/a + \frac{1}{2}$; normalized coordinate transverse to a plane Clogston 2 line (Section XII).
ρ :	Radial coordinate in cylindrical system.
ρ_1, ρ_2 :	Inner and outer radii of main dielectric in coaxial Clogston line.
σ :	$\sqrt{i\omega\mu(g + i\omega\epsilon)}$; intrinsic propagation constant of medium.
ϕ :	Angular coordinate in cylindrical system. Phase angle, $\tan^{-1}(\epsilon''/\epsilon')$, of complex dielectric constant.
χ :	$-i\Gamma_\ell$; a transverse separation constant.
ω :	Angular frequency in radians·second ⁻¹ .

FUNCTION SYMBOLS

Re:	Real part.
Im:	Imaginary part.
log:	Natural logarithm.
sh:	Hyperbolic sine.
ch:	Hyperbolic cosine.
J_0, J_1 :	Bessel functions of the first kind.
N_0, N_1 :	Bessel (Neumann) functions of the second kind.
I_0, I_1 :	Modified Bessel functions of the first kind.
K_0, K_1 :	Modified Bessel functions of the second kind.

Electrical Noise In Semiconductors

By H. C. MONTGOMERY

(Manuscript received June 3, 1952)

Transistors, diodes, and single crystal filaments of germanium have common noise properties: a spectrum varying inversely with frequency, and strong dependence on the biasing current. Theoretical attempts to explain this noise are reviewed briefly. Experiments with single crystal filaments indicate that the noise resides in the behavior of the minority carrier. In one type of experiment, the correlation of noise voltages in adjacent portions of a filament is quantitatively related to the lifetime and transit time of minority carrier. In another, the effect of a magnetic field on the noise is found in accord with calculated changes in lifetime of the minority carrier.

In the development of the transistor it was recognized quite early that electrical noise in the device was considerably in excess of Johnson noise, particularly at low frequencies. Noise having a similar spectrum had been observed many years earlier in microphonic carbon contacts carrying a current, and in copper oxide rectifiers, composition resistors, and crystal diodes. Flicker noise in vacuum tubes appears to be a related phenomenon. A number of attempts have been made to determine the mechanism of production of noise of this sort, but none have been particularly successful.

In this paper we will first survey the more important characteristics of noise in germanium diodes and transistors. This will be followed by a partial hypothesis as to the nature of the noise mechanism. We will then discuss experimental work on noise in filaments of single crystal germanium carrying a dc current. These experiments strongly support the hypothesis, and in fact led to its formulation in the first place.

I. NOISE IN DIODES AND TRANSISTORS

There are many similarities in the noise phenomena found in diodes and transistors of both the point contact and junction type. It seems likely that the noise mechanism is similar in all these devices.

One of the most characteristic features of the noise in such structures

is the spectrum. The spectral density (power per unit bandwidth) varies inversely as the frequency, according to the relation

$$dW = f^{-n} df$$

where the exponent lies between 1 and 1.5 with an average about 1.2. This type of spectrum will be referred to as a $1/f$ spectrum. Measurements of the spectra of silicon point contact diodes have been reported by P. H. Miller¹ for the frequency range 20 cycles to 300 kilocycles. Spectra of point contact transistors measured by the author have been reported elsewhere^{2, 3} for the range 20 to 15,000 cycles. Typical spectra for p - n junction type diodes and transistors are shown in Fig. 1. Almost without exception, our measurements and those reported in the literature have shown the $1/f$ spectrum over most of the frequency range covered. There is some evidence from the related fields of flicker noise and carbon microphone noise that the $1/f$ spectrum may extend to frequencies well below 0.1 cycle per second. Some departures from this type of spectrum have been noted in the neighborhood of 100 kc, as shown in the curves.

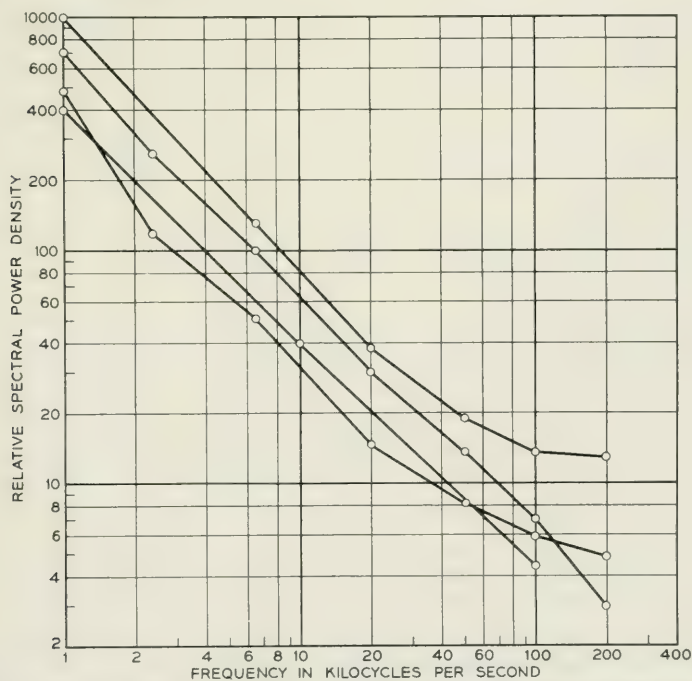


Fig. 1—The spectrum of noise in n - p - n transistors varies inversely with frequency.

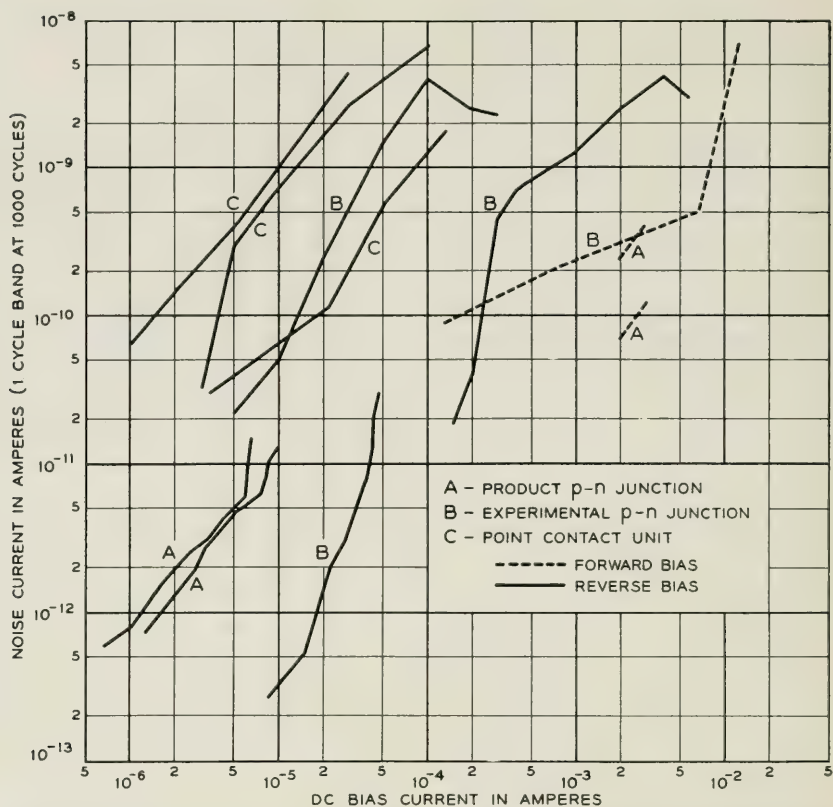


Fig. 2—The short-circuit noise current from a point contact or junction diode generally increases with dc bias current.

A second characteristic feature of noise in all semiconductor devices is that it is current dependent. In the absence of biasing current only Johnson noise is observed. When biasing current is present the noise power may be as much as three or four orders of magnitude above Johnson noise. As a general thing the noise increases as the bias is increased, although some minor exceptions to this rule are noted, usually at bias values where the slope of the current-voltage curve is changing rapidly.

To illustrate the bias-dependent behavior, the noise properties of some germanium diodes of various types are shown in Fig. 2. The short circuit noise current in a 1-cycle band at 1000 cycles is plotted as a function of dc bias current, some of the data being for forward bias, but most for reverse bias. Several curves are shown for each type of unit, and

these are typical of the variations encountered. There is a general tendency for noise current to increase in proportion to bias current, but in limited regions the individual units may have slopes considerably different from unity. It would perhaps be more logical to plot current densities rather than total currents, but because of the general form of the relations this makes little difference in the overall picture, and there is some difficulty in estimating the appropriate area for the point contact units. There is an almost unlimited number of different ways of representing noise data. For example, noise current, current density, voltage, or available power may be expressed as a function of various bias parameters. Of a good many combinations tried, none gave an outstandingly simple picture of noise behavior, and the representation used in Fig. 2 is probably as good as any for an overall picture of diode noise.

The noise behavior of transistors depends on two bias parameters. Selection of the emitter current and collector voltage for the parameters usually leads to a rather simple representation. It often turns out that the noise behavior as an amplifier over the commonly used range of bias values depends largely on the collector voltage and is relatively independent of the emitter bias. Data of this sort were shown for point contact transistors in a previous reference,³ and have been given for an n - p - n transistor by Wallace and Pietenpol.⁴ A somewhat more complete family of curves is shown in Fig. 3 for a recent n - p - n transistor.

A few attempts have been made to determine the effect of tempera-

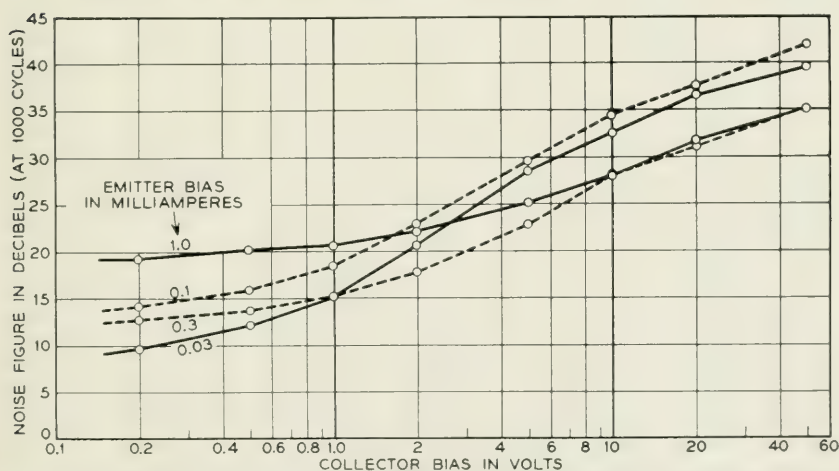


Fig. 3—The noise figure of an n - p - n transistor depends in a fairly simple way on emitter current and collector voltage.

ture on noise behavior. Such experiments have been rather unsatisfactory because the changes in impedance and gain characteristics as a function of frequency are of the same order as the changes in noise properties. This makes the interpretation ambiguous. By and large, such experiments suggest that changes in noise with temperature are rather small, perhaps of the order of the change in absolute temperature, and not at all like the exponential changes associated with a diffusion process. This observation does not necessarily rule out a diffusion-like noise process; it might indicate merely that we are not looking at the right part of the spectrum to observe exponential changes with temperature.

II. A HYPOTHESIS REGARDING THE NOISE MECHANISM

Considerable work has been done on the theory of current-dependent noise having a $1/f$ spectrum. Among the earliest was that of Schottky⁵ in connection with flicker noise in vacuum tubes. He considered the arrival of foreign atoms on the emitting surface of the cathode as a random series of events governed by a diffusion law with a characteristic time constant, and arrived at a $1/f^2$ rather than a $1/f$ spectrum, and a highly temperature sensitive process. Surdin⁶ pointed out that by postulating a series of decay processes with suitably distributed time constants a $1/f$ spectrum could be achieved. From physical arguments regarding the emission process from cathodes, Macfarlane⁷ obtained a range of relaxation times and a $1/f$ spectrum, in a process which was highly temperature dependent. Richardson⁸ gave a very general analysis of the noise properties of systems in which the conductivity was governed by a diffusion process. One conclusion was that a geometrically simple diffusion process in one, two, or three dimensions could not lead to $1/f$ spectrum, although by some highly specialized assumptions about the geometry of a contact surface he was able to obtain such a spectrum. DuPré,⁹ in considering a hypothesis somewhat resembling that of Surdin, showed that the required range of activation energies was physically reasonable, and that the assumptions could be set up in such a way as to make the process relatively temperature independent. Several of the above authors and Van der Ziel¹⁰ discuss the physical basis for applying flicker noise theory to the noise in semiconductors. Although this theoretical work has contributed a great deal to distinguishing between suitable and unsuitable mechanisms, there is still no specific physical theory of noise in semiconductors which can be tied in a quantitative manner to experimental results.

The experimental work described in the remainder of this paper has

led to a hypothesis regarding the noise mechanism, which is by no means a complete explanation, but which may be a useful step in that direction. This hypothesis resulted largely from the experimental work, but it seems worth while to describe it first to help appreciate the significance of some of the experimental results.

It has been observed that in many semiconductor structures the noise voltage is approximately proportional to the dc bias current. This relation suggests that the noise is the result of fluctuations of the conductivity of the material, which modulate the bias current and produce a fluctuating voltage across the specimen. Such fluctuations in conductivity could result from variations in concentration of the minority carrier (holes in *n*-type material, electrons in *p*-type). The magnitude of the observed noise and the type of spectrum seem to demand that the fluctuation be coarse-grained in time to a much greater extent than could be accounted for by random statistical fluctuations of carrier density. Experiments of Haynes¹¹ on lifetime and transit of injected carriers in rods of germanium have occasionally indicated finite sources of minority carriers in the material. Our hypothesis is that such sources of carriers are rather generally distributed over the material (although mostly too small to be noticed in experiments of the Haynes type), and that their activity is being modified at a slow rate by some unspecified local influence in a suitable way to agree with the observed noise spectrum.

The experiments described below involving noise correlation phenomena and the effect of a magnetic field on noise point strongly to an important role for the minority carrier in the noise mechanism, and hence strongly suggest some such hypothesis as that just described.

III. NOISE IN SINGLE CRYSTAL FILAMENTS

It was found several years ago that a filament cut from single crystal germanium of high purity exhibits noise well above Johnson noise when a dc current is flowing in it. It is not clear whether this noise arises in the body of the material or on the surface, but to date no method of preparing the sample has eliminated this noise, and it is a prominent feature even at bias fields as low as 10 volts per centimeter. This noise seems to have most of the characteristics of the noise in diodes and transistors: it has the $1/f$ spectrum, is current dependent, and is quite stable with time. It has been the subject of considerable study in the hope that a better understanding of it would illuminate the whole field of semiconductor noise.

Samples, referred to as "bridges", have been cut from thin slabs of single crystal germanium, by a technique devised by W. L. Bond,¹² often of a form shown in Fig. 4. Side arms for both the current and the noise measuring electrodes have been found necessary to avoid spurious noise at the electrodes. A large inductance in the bias circuit greatly reduces the effect of any noise voltage generated at the bias electrodes. The spurious noise power from this source is seldom more than a few per cent of that being measured. It should be noted that the contact area for the noise measuring electrodes should not be on a portion of the specimen carrying bias current, otherwise spurious noise may be generated at these electrodes. Typical dimensions for the straight central filament of the bridge are $0.05 \times 0.05 \times 0.7$ cm. The side arms have sandblasted surfaces to suppress holes or electrons injected at the electrodes. The central portion may be etched, sandblasted, or otherwise treated at will. The enlarged circular areas are rhodium plated to provide good contacts to each side arm.

Measurements of the noise spectrum in such bridges with several different etching treatments and with sandblasted surfaces are characterized by the $1/f$ spectrum over a wide frequency range.* Fairly extensive measurements have been made in the audio frequency range, and a few covering the range from 20 cycles to 1 megacycle. A typical spectrum is shown in Fig. 5.

The current dependence of the noise is shown in Fig. 6 for a number of samples, mostly n -type, one p -type, and with various resistivities. The outstanding feature is that noise voltage always increases with dc bias voltage. In many cases there is direct proportionality at the lower bias values, increasing to a square law at higher biases. There are some

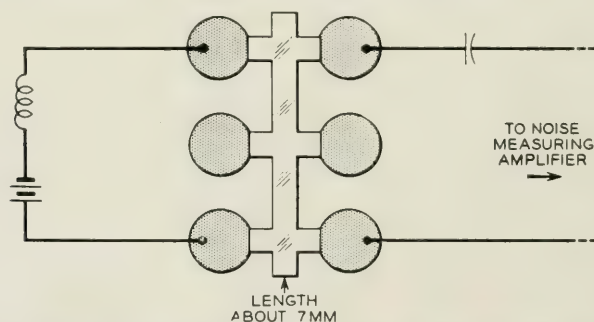


Fig. 4—Filament with side arms cut out of a single crystal of germanium.

* Departures from the $1/f$ spectrum at frequencies of the order of 100 kilocycles and above were first discovered by G. B. Herzog and A. Van der Ziel. See Reference 13.

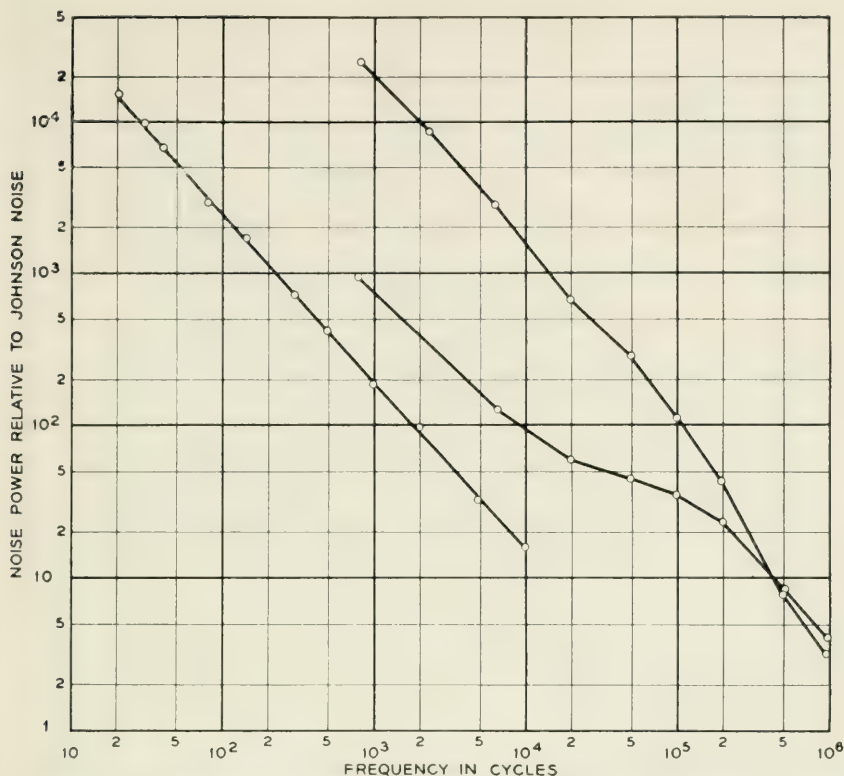


Fig. 5—Typical spectra of noise in single crystal filaments carrying a dc current.

exceptions to this trend. Also, there are large variations in the magnitude of the noise. An average unit shows a noise voltage about three times Johnson noise at a bias of 10 volts per centimeter.

The noise behavior at reduced temperatures has been investigated. Results on three different bridges are shown in Fig. 7. The open circuit noise voltage is shown as a function of temperature for constant bias voltage. Although the curves show rather large irregularities, there seems to be no general trend for noise to decrease with decreasing temperature over the range covered, from -200°C to room temperature.

The surface treatment applied to a bridge may affect the noise very substantially. A sandblasted surface usually gives the lowest noise. Etching the surface may raise the noise voltage by a factor of ten or more, though the dc resistance changes only a few per cent. The technique of washing and drying the surface may have an important effect

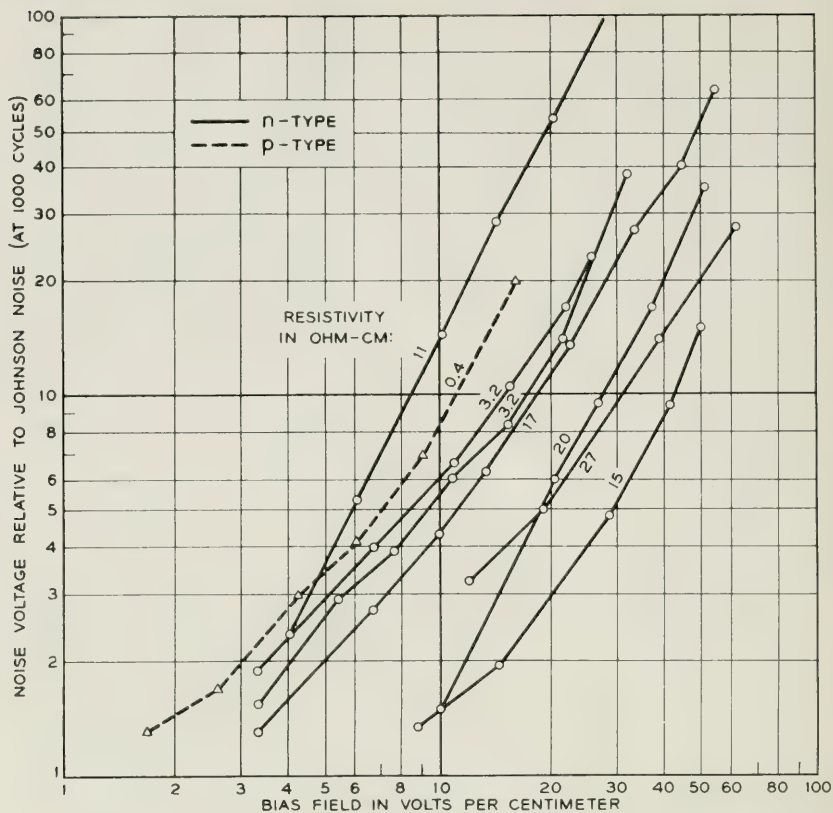


Fig. 6—Variation of noise with dc bias in single crystal filaments.

on the noise. Some of these processes also affect the lifetime of carriers in the bridge to a large extent. However, there seems to be no direct and simple relation between the two effects, since treatments have been found which change the noise by a large factor with almost no effect on lifetime, and vice versa.

Fig. 8 shows measurements of noise voltage on several dozen bridges at a uniform bias of 10 volts per centimeter, all having sandblasted surfaces, mostly of *n*-type but a few of *p*-type germanium, and with widely different values of resistivity, produced by varying impurity concentrations. There is considerable scatter in the results, but there is a fairly obvious tendency for noise voltage to increase in proportion to resistivity. Since Johnson noise also increases in proportion to resistivity in a structure of fixed dimensions, the conclusion is that with constant bias voltage the ratio of current induced noise to Johnson noise tends to be

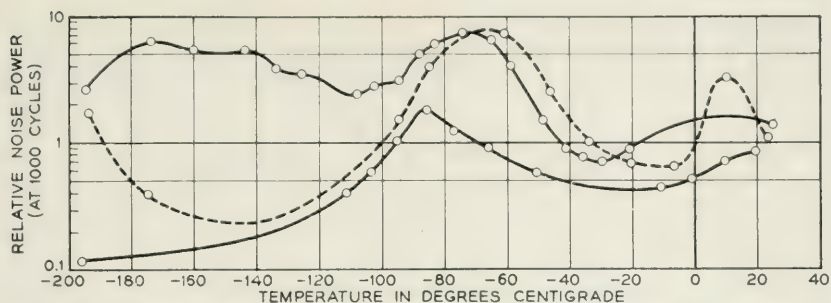


Fig. 7—Variation of noise with temperature in single crystal filaments.

independent of the resistivity of the material. From the data it also appears that there is no consistent difference between n - and p -type material.

Noise does not appear to depend on orientation of the filament with respect to the crystal axes. Filaments orientated along the 100, 110, and 111 directions and rotated in several ways about these directions showed

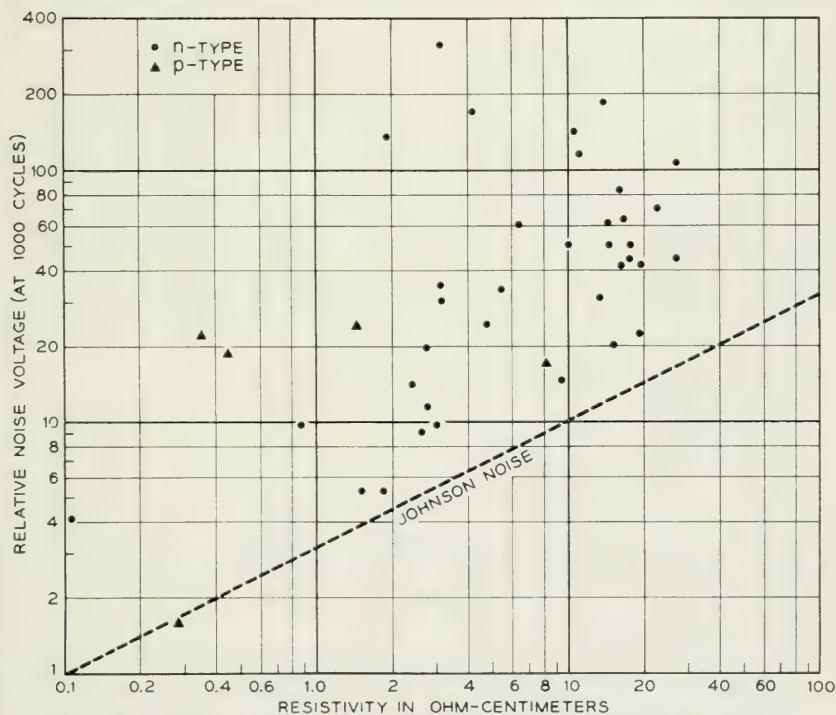


Fig. 8—Variation of noise with resistivity in single crystal filaments.

no significant differences in noise behavior. It should be noted, however, that small variations might be hidden in the large scatter in the data from undetermined causes.

IV. NOISE AND MAGNETIC FIELDS

An important role for the minority carrier in the noise mechanism was first clearly indicated in experiments on the effect of a magnetic field on noise in germanium filaments. It has been found experimentally that the noise in a single crystal filament may change by a substantial factor when the filament is subjected to a steady transverse magnetic field. The following discussion will show that this behavior is in harmony with the hypothesis of noisy injection of minority carriers, as set forth in a preceding section.*

The physical picture on which this treatment is based involves the random injection of holes into an n -type filament by hole sources which may be either in the interior or on the surface of the filament.† It is assumed that the spectrum of the noise arises from the fluctuating nature of the noise source. The effect which any source has will depend upon the lifetime of the holes which it emits. If these holes remain in the filament for a long time, they will produce more noise than if they remain in the filament for a short time. We shall be concerned with the effect of magnetic fields upon these lengths of time and shall not deal in this paper with the fluctuations of the noise sources themselves. If a transverse magnetic field is applied to an n -type germanium filament, a Hall effect voltage is set up and the holes will be deflected towards one surface of the filament. Since recombination takes place principally at the surfaces, this may cause a substantial change in the lifetime of the holes. In order to determine the effect of the magnetic field on the noise we proceed along the following lines.

(a) We assume that the observed noise is due to fluctuations in the conductivity of the filament produced by fluctuations in the hole concentration. Since these fluctuations are small, we may take the change in conductivity to be proportional to the change in average hole den-

* The following semi-quantitative theory of the dependence of noise on magnetic field is taken with some modification from unpublished work of W. Shockley and H. Suhl, on the basis of which the calculations leading to the curves of Figs. 10 and 11 were carried out. It is hoped that this work may be published in the near future.

† To simplify the terminology, the discussion is based on n -type material with holes as minority carrier. An exactly similar argument could be made for p -type material with electrons as the minority carrier. There is some experimental evidence of the similarity of behavior of n - and p -type germanium, though most of the experimental work has been done with n -type.

sity. (b) We restrict the noise measurements to frequencies low enough so that the period is long compared to the lifetime of a hole. It is then evident that the contribution of a hole source to the noise is proportional to the fluctuating hole current generated by the source and to the average lifetime of the holes. This lifetime depends on the position of the source in the filament, the absorption properties of the surfaces and the electric and magnetic fields.* (c) We assume that the generation properties of the sources are unaffected by the magnetic field, hence, the calculation of the effect of the field on the noise reduces to a problem of calculating the change in lifetime produced by the field. (d) We neglect body recombination in comparison with surface recombination. In germanium filaments of the size usually dealt with, this approximation causes only a small error in the lifetime. (e) Individual sources (or at any rate groups of sources over regions small compared to the dimensions of the filament) will be considered to be statistically independent; therefore, the total effect on the noise can be determined by summing the squares of the contributions from individual sources. Hence we wish to evaluate the following expression:

$$\text{Change in noise power at field } H = \langle \tau^2(H) \rangle / \langle \tau^2(0) \rangle \quad (1)$$

where the symbol $\langle \rangle$ indicates an average over all the noise sources. The statements (a) to (e) represent the principal assumptions in developing the theory.

In order to calculate τ as a function of the magnetic field, H , we consider a steady state case in which a current of holes J_0 is injected into a region in which the average lifetime is τ . If the density of holes in the region is $p(x, y, z)$, the total number is

$$P = \int p(x, y, z) \, dx \, dy \, dz.$$

However, $P = J_0\tau/q$, where q is the charge carried by a hole. Therefore,

$$\tau = \frac{q}{J_0} \int p(x, y, z) \, dx \, dy \, dz \quad (2)$$

This is the method of evaluating τ which is used in the qualitative discussion which follows, and also in the calculation of the curves of Figs. 10 and 11.

* It should be pointed out that a consequence of the hole injection theory of noise in a filament is that marked frequency dispersion should occur when the frequency being studied is high enough so that a period is short compared to the lifetime of holes in the filament. However, we shall neglect this important and interesting aspect of the problem.

Three cases will be treated. In all of these it will be supposed that the width of the filament parallel to the magnetic field is relatively large, so that effects from the edges can be neglected. Also, we are concerned only with average effects over the long dimension of the filament. This permits us to deal with a one-dimensional problem. We shall consider first the case in which holes are supposed to be injected from the surfaces, and the two surfaces have equal and rather large recombination rates. In Fig. 9, part (a) shows how holes injected from each surface are distributed across the thickness in the absence of a magnetic field, and part (b) shows the distribution with a moderate field. The form of these distributions may be determined from the following arguments.

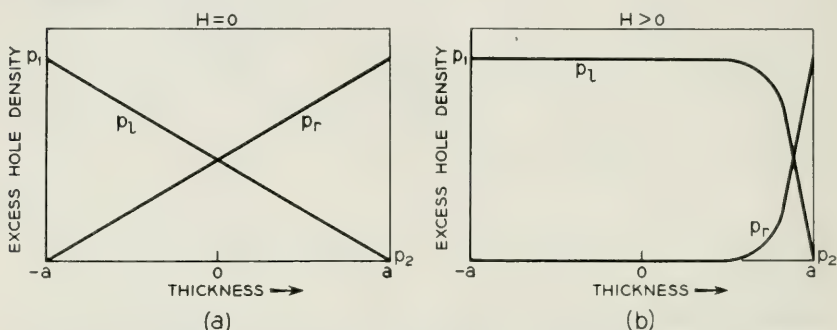


Fig. 9—Excess hole density across the thickness dimension, (a) with no magnetic field, (b) with moderate magnetic field.

If we suppose a steady hole current J_0 emitted from the left-hand surface of the filament, then a relatively high concentration p_1 of holes will appear directly in front of the surface. Some of these holes will recombine upon the surface, the rate J_1 being given by

$$J_1 = p_1 S q$$

where S is the recombination constant for the surface. The balance of the holes will diffuse through the filament to recombine upon the right surface at a rate

$$J_2 = p_2 S q$$

and we note that $J_1 + J_2 = J$. Because of the high recombination rate, p_2 will be very small; hence, J_2 will be much smaller than J_1 . In the absence of a magnetic field the gradient is uniform, and the concentrations will be linear, as shown in part (a) of the figure. An identical argument

applies to p_r , the concentration of holes emitted from the right-hand side.

Under the influence of a magnetic field pushing holes toward the right, the concentrations will change to those shown in part (b) of the figure. The magnetic field will pull holes through the filament and tend to prevent diffusion from right to left. For some moderate value of field, the value of J_2 is not increased enough to change J_1 appreciably, so the value of p_1 is nearly the same as with no field. At the same time the effects of diffusion are suppressed by the field so that the concentration p_1 extends nearly to the right side of the figure. By the same action, the concentration of holes emitted from the right surface drops to zero very quickly.

From the curves of Fig. 2 and relation (2), we see that the area under the density curve, and hence the lifetime of holes injected at the left is at most doubled by the magnetic field, while the lifetime of those injected at the right is reduced nearly to zero. Recalling that the noise is proportional to a summation of the square of the lifetimes, we see that the noise power is at most doubled at a suitable value of magnetic field.

Higher values of field will sweep so many holes to the right-hand surface as to substantially reduce p_1 , so at very high fields the noise decreases monotonically to zero.

Thus it is seen that the noise behavior is the result of competing tendencies. On the one hand, the magnetic field helps holes escape from the surface at which they are emitted, but on the other hand it tends to push these holes against the opposite surface and thereby reduce their lifetime. The relative importance of those two tendencies depends on the surface recombination properties and the strength of the magnetic field.

Calculation of the lifetime along the lines just discussed involves solution of the continuity equation

$$D \frac{d^2 p}{dx^2} - \mu_p E_H \frac{dp}{dx} = 0$$

with suitable boundary conditions. The results of such a calculation carried out by Shockley and Suhl in the work already referred to are plotted in Fig. 10.

In order to make the results independent of sample dimensions, the following parameters are used. The first parameter is proportional to the applied magnetic field, and is defined as the effective transverse potential in units of kT/q :

$$\Phi = \frac{tE_H}{kT/q} = 172tEH \times 10^{-5} \quad (3)$$

where t = thickness of the filament (cm)

H = magnetic field (oersteds)

E = applied electric field (volts per cm)

E_H = effective transverse component due to Hall effect (volts per cm)

q = unit electronic charge

kT = Boltzman's constant \times absolute temperature.

The constant may be derived by noting that kT/q is 1/40 volt at room temperature, and that the effective transverse field, E_H , may be expressed as follows. (See Reference 14, Section 8.8.)

$$\begin{aligned} E_H &= \theta E = (\theta_n + \theta_p)E \\ &= (\mu_n + \mu_p)HE \times 10^{-8} \\ &= 4.3 \times 10^{-5}HE \end{aligned}$$

where θ = Hall angle

μ_n = Hall mobility for electrons (2800 cm²/volt-sec)

μ_p = Hall mobility for holes (1500 cm²/volt-sec).

The other dimensionless parameter is proportional to the rate of surface recombination, and is defined as the ratio of the surface recombination velocity to the diffusion velocity from the center:

$$\psi = st/2D = st/86$$

where s = recombination velocity characteristic of the surface (cm/sec)

D = diffusion constant (cm²/sec).

The numerical constant is given for holes at room temperature. The noise changes are expressed in decibels, that is, ten times the common logarithm of the ratio of noise powers with and without the magnetic field.

A second case is that in which generation and recombination are on the surfaces, but the two surfaces have unequal absorption properties. It might be expected that rather large increases in noise would result when the magnetic field was poled to pull holes away from the surface with high absorption properties, and this turns out to be the case when the calculations are carried out. The results are shown in Fig. 11 for a

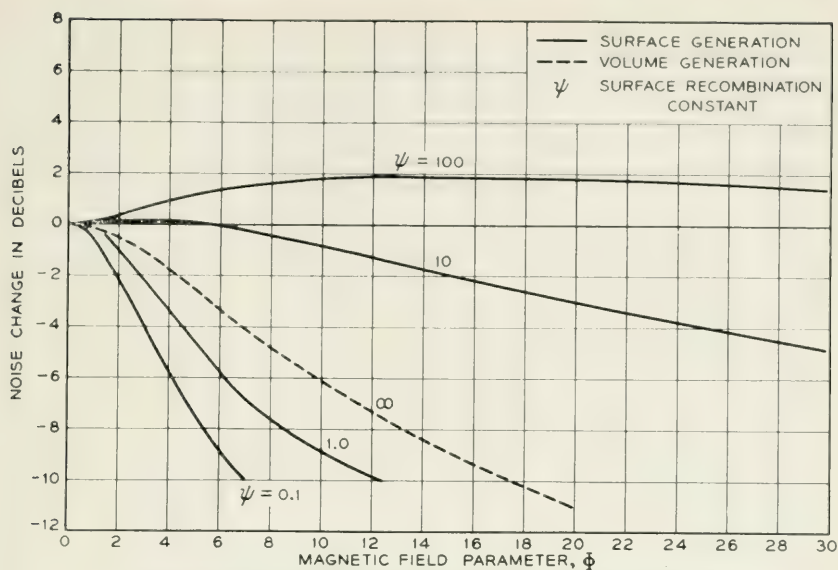


Fig. 10—Calculated magnetic effect for similar surfaces.

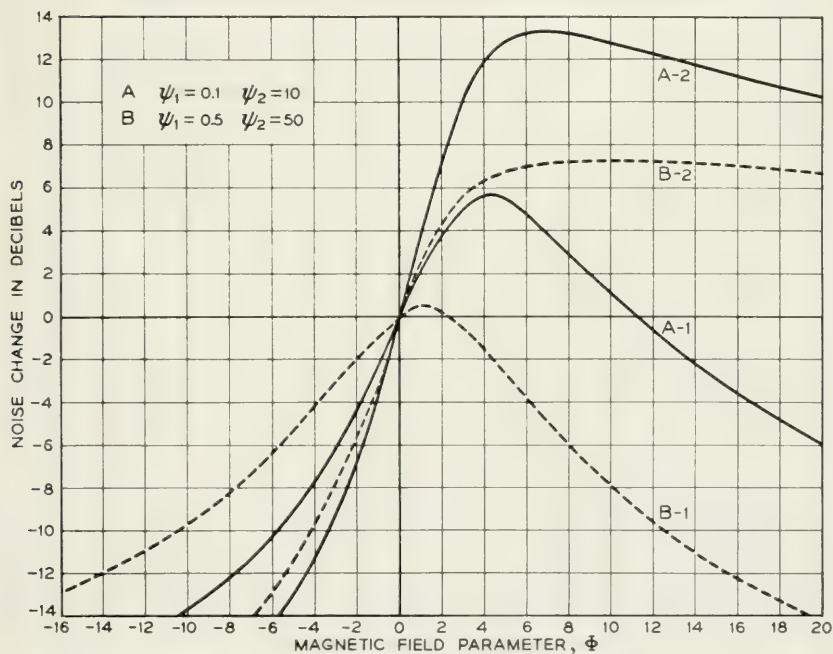


Fig. 11—Calculated magnetic effect for dissimilar surfaces. Curve 1 of each pair is for the contribution from the surface having the lower recombination constant.

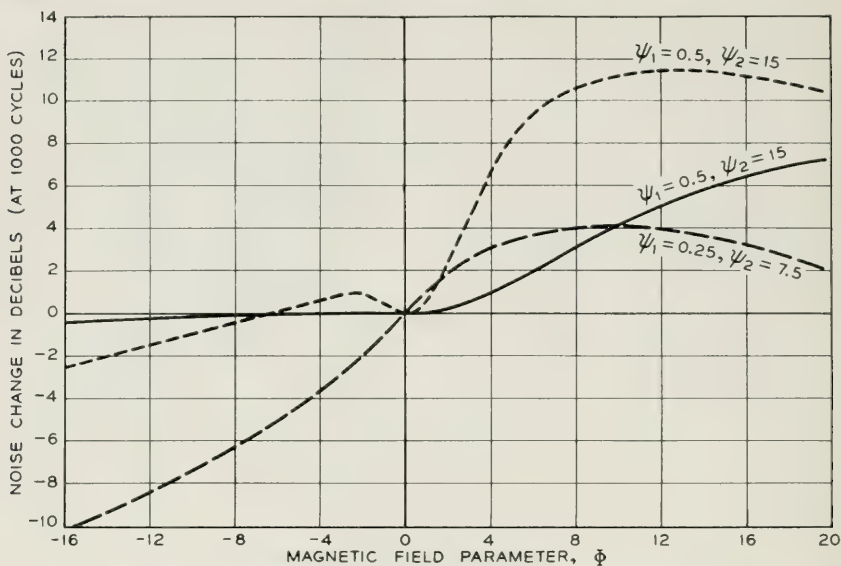


Fig. 12—Experimental magnetic effect for dissimilar surfaces.

case where the two recombination parameters are 0.1 and 10, and for a second case where the parameters are 0.5 and 50. In this figure, separate curves have been shown for noise due to holes generated on each of the two surfaces. The total noise would be gotten by adding the noise powers represented by the two curves after appropriate weighting for the contributions of the two surfaces. At present we do not see any way of determining the weighting factor.

A third case is that in which it is assumed that the noisy generation of holes is uniform throughout the body of the filament, but that recombination takes place on the surfaces only. These assumptions seem at first sight to be in contradiction to the statistical mechanical principle of detailed balancing, which states that under equilibrium conditions all processes occur with equal frequency in the forward and reverse directions. Thus it would seem that if holes are generated in the interior, we must consider recombination in the interior also. Actually this is not necessary under the non-equilibrium conditions which prevail during noise measurements. There is no necessity for the noise generated by a source and a sink for holes to be simply related to the strength of this source. Thus we may suppose there are relatively weak sources and sinks for holes in the interior, but that the hole absorption and generation of the sources is very noisy compared to the recombination and generation processes on the surfaces. If this is the state of affairs, most

of the noise will be generated in the interior, but a hole generated in the interior will be much more likely to recombine on the surface. The dotted curve of Fig. 10 has been calculated assuming a uniform distribution of noise sources throughout the interior of the filament and equal and very large recombination constants for the two surfaces. It is seen that for this case the reduction of lifetime predominates, and there is a monotonic decrease in noise with increasing magnetic field.

Experimental work has given results which in most cases are in fair qualitative agreement with the calculated relations. Measurements for three filaments, each of which had one high recombination and one low recombination surface, are shown in Fig. 12. The recombination parameters, as shown on the curves, were of the order of $\psi = 10$ for one surface, and $\psi = 0.5$ for the other. The general shape of the curves is quite similar to the calculated curves of Fig. 11. The maxima are of the right order of magnitude, and occur at reasonable values of the field parameter Φ . The lack of detailed agreement between the measured and calculated curves is not surprising, because the experimental conditions did not fulfill the assumptions made for the calculations in several respects. The filaments were neither wide enough nor long enough so that edge and end effects could be overlooked. The recombination properties of the surfaces could not be measured directly, but had to be estimated from other filaments which had been similarly treated. One experimental curve shows a secondary maximum on the opposite side of the origin. This might indicate a defective portion of one surface having an anomalous recombination constant.

Experimental results are shown in Fig. 13 for four filaments, each of which had nominally equal recombination constants for the two surfaces. These may be compared with the calculated curves of Fig. 10. It will be noted that the experimental curves are not symmetrical about $\Phi = 0$. This lack of symmetry is probably due to dissymmetry in the

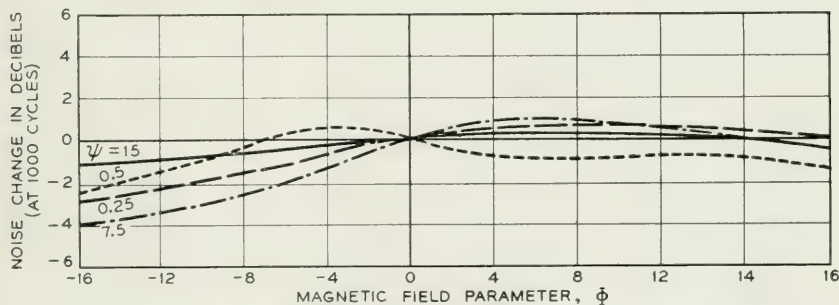


Fig. 13—Experimental magnetic effect for similar surfaces.

samples, particularly the fact that one surface of each filament was cemented to a support, which would probably change the surface recombination properties somewhat. Aside from the lack of symmetry, the behavior of the two filaments with the higher recombination constants is in reasonable agreement with the calculated curves. The filaments with the lower recombination constants are in poor agreement with calculated values, in that the noise does not fall off with increasing field nearly as fast as calculated. The cause of this behavior is not understood. These experimental curves may also be compared with the dotted curve of Fig. 10, calculated on the assumption of volume generation and surface recombination. The similarity is quite poor in all cases. The somewhat better agreement with the surface generation calculations than with the volume generation calculations is not the basis for anything more than a very tentative feeling that the experimental results support the surface generation viewpoint.

While there are many discrepancies in detail between the experimental and calculated relations between noise and magnetic field, these are at least partially understandable in terms of the differences between the experimental setup and the theoretical model. The high degree of qualitative agreement considerably strengthens the hypothesis of noisy injection of minority carriers as an important element in the noise process.

V. NOISE CORRELATION PHENOMENA

The noisy hole injection hypothesis leads one to expect certain correlation phenomena in the noise voltage observed in neighboring portions of a filament. Consider first noise measurements at a frequency so low that the transit time of a hole* along the filament is negligibly small. This might be a frequency of one kilocycle in a typical experiment. The holes have an average lifetime, from which can be determined an average life path, which is defined as the product of the lifetime by the drift velocity under the existing electric field. Noise voltage measurements across segments of the filament much shorter than a life path should be highly correlated, since nearly all the holes which make a transit of one segment will make an almost simultaneous transit of the other segment. On the other hand, noise voltages across segments much longer than a life path should show little correlation, because most of the holes appearing in the two segments are from different sources, and the sources have been assumed to be statistically independent.

* As before, the concepts apply equally well to electrons in *p*-type material.

A second situation arises when noise is measured at frequencies high enough so that the transit time of holes between segments is not negligible. In this case we should expect the correlation between the noise voltages to be improved by incorporating in one channel of the measuring circuit a delay equal to the transit time between segments.

In order to calculate the correlation resulting from the first situation, we set up a theoretical model based on a few simplifying assumptions: (a) The noise process may be represented by an array of noisy hole current generators which are statistically independent; (b) These generators are uniformly distributed along the filament over the segments where the noise is to be observed, and for a sufficient distance on either side to produce uniform conditions over the segments; (c) The hole currents from the generators decay exponentially with a decay constant determinable from the lifetime; (d) Measurements are made at low enough frequencies so that time of transit of holes may be neglected. We will consider later an alternative to the second assumption. The correlation coefficient between two voltages of instantaneous values v_1 and v_2 may be defined as

$$\rho_{12} = \overline{v_1 v_2} / (\overline{v_1^2} \times \overline{v_2^2})^{1/2}$$

where the bars represent time averages. To evaluate this expression, the contribution of a single generator to the noise voltage in each segment is determined by integrating over the appropriate portion of the decay curve. The total contribution from all generators to the mean voltage product and the mean squared voltages is then determined by integrating the product or square over all the generators. The details are carried out in the appendix, and lead to the solid curve of Figs. 14-16, in which the ordinates are the correlation between noise voltages in two segments of a filament and the abscissae are the ratio of life path of a hole to the segment length.

In an experiment the lifetime τ of holes remains fixed, determined chiefly by the recombination properties of the surface. Consequently the life path ℓ is proportional to the hole velocity, which is determined by the electric field, according to the relation

$$\ell = \tau \mu E$$

where E is the applied field in volts per centimeter and μ is the drift mobility of holes. Hence, by varying the biasing voltage a large range of life path values can be obtained.

The correlation is measured by carrying the noise voltages through separate amplifying channels having identical pass bands extending

from 800 to 1300 cycles per second. A switching arrangement makes it possible to apply either of the output voltages or their sum or difference to a rectifier-meter combination. From the readings of the meter the correlation can be computed according to the relation

$$\rho_{12} = (S^2 - D^2)/4V_1V_2. \quad (6)$$

V_1 and V_2 are rms values of the individual noise voltages, and S and D are the rms values of their sum and difference. The equivalence to expression (5) can be seen by noting that

$$S^2 - D^2 = (\overline{v_1 + v_2})^2 - (\overline{v_1 - v_2})^2 = 4\overline{v_1 v_2}.$$

Results of correlation measurements on three bridges are shown in Figs. 14–16. In each case the calculated curve is shown for reference. The values of ℓ were calculated from decay measurements on optically injected holes, as described by J. R. Haynes,¹¹ using a value for mobility of $1700 \text{ cm}^2/\text{volt-sec}$. In Fig. 14 the agreement with the theoretical model is very good. The scatter in the points is due to fluctuations in the noise, which are quite large in the band used for these measurements. In Fig. 15 the agreement could be made quite good with a lateral shift

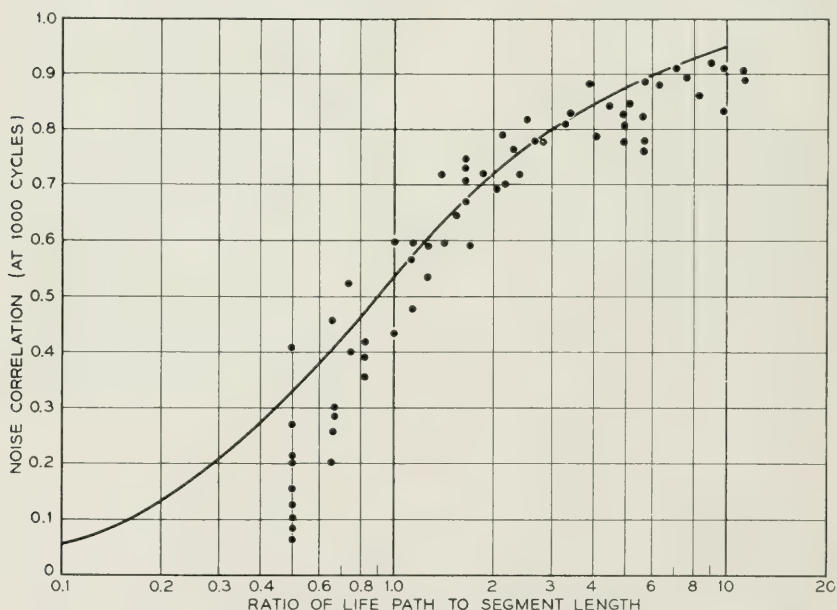


Fig. 14—Noise correlation. The solid curve is calculated, the points experimental.

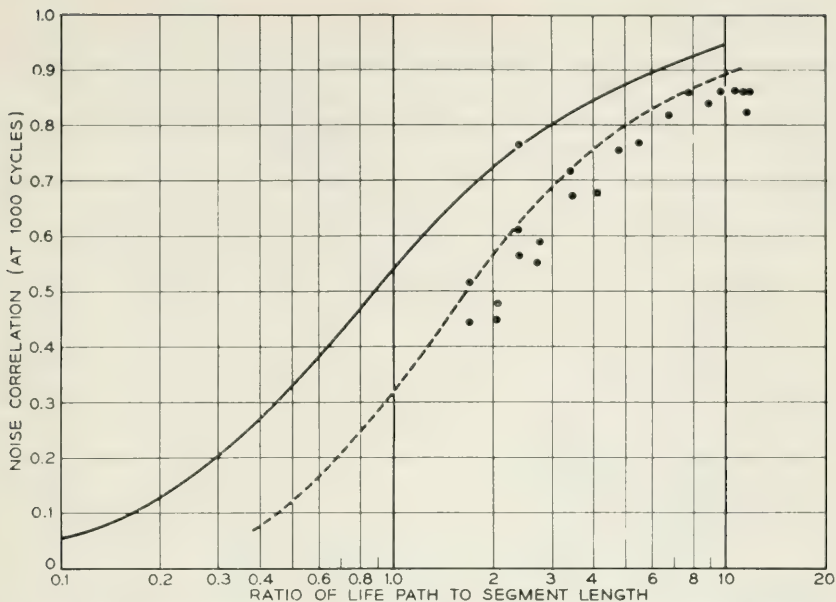


Fig. 15—Noise correlation. The dotted curve includes allowance for losses at the side arms.

by a factor of two. In Fig. 16 the form of the experimental curve seems different from that calculated. In particular, the slope is steeper, and the curve tends to level off at a correlation of about 0.8. It seems possible to explain the discrepancies between the experimental data and the calculations on the basis of two considerations which were not included in the model. (a) The pair of side arms separating the two segments of the filament serve to drain off some holes which would otherwise contribute to the correlation. The dashed curve in Fig. 15 shows the calculated effect, on the assumption that the absorption in the side arms is equivalent to an extra section of filament equal in length to half a segment. The actual distance across the side arms is only 20 per cent of a segment, but it is not hard to believe that the decay rate in this region might increase by a factor of two or three due to the reduced electric field and loss of holes down the side arms. (b) The model assumed a uniform distribution of noise sources along the filament. There is experimental evidence that the distribution may be quite spotty. This can have a substantial effect on the form of the correlation curve. For example, the dashed curve in Fig. 16 shows the curve calculated for noise sources lumped at the mid-point of each segment. Other assumed positions might shift the curve considerably along the horizontal axis.

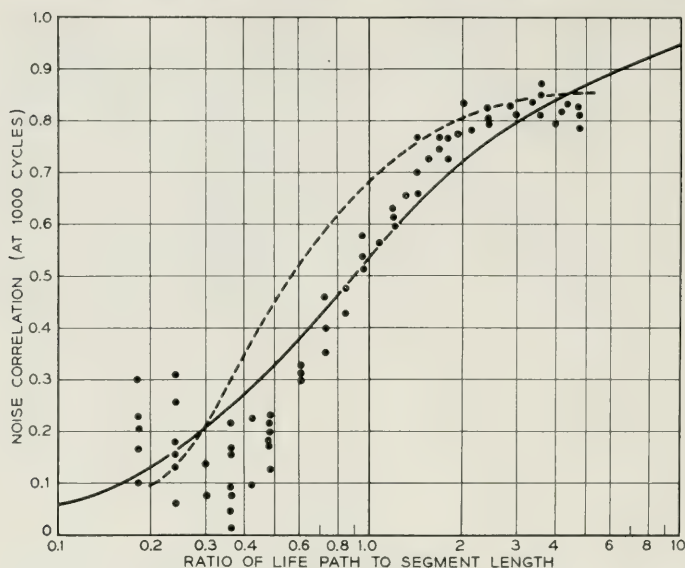


Fig. 16—Noise correlation. The dotted curve is calculated for lumped noise sources.

In view of these considerations there seems to be very satisfactory agreement between the experimental results and the model.

Another type of experiment involves noise measurements at frequencies high enough so that the transit time of a hole across a segment is an appreciable fraction of a cycle. In this case the correlation between noise voltages from adjacent segments can be improved by putting a time delay in one channel of the measuring circuit. Measurements were made by taking the noise voltages from the two segments through separate amplifying channels having identical pass bands extending from 17 to 24 kilocycles. The outputs of the two channels were put on the vertical and horizontal plates of a cathode ray oscilloscope, forming a sort of Lissajous pattern. The patterns differ from those obtained with sinusoidal voltages in that the elliptical figures are filled in solid, due to the continual variation in amplitude of the noise. A phase shifting device is included in one channel, and as the phase is shifted to give optimum correlation, the elliptical pattern narrows down and approaches a line inclined at 45° . For a quadrature phase shift, the pattern becomes circular, and in practice this setting can be determined more precisely than the in-phase setting, largely because the background noise in the circuit is less troublesome. With the phase shift for optimum correlation determined, the delay at the center of the pass band is easily calculated,

and since the band is not very wide, the variation in delay over the band is not important. From the drift mobility of holes we may estimate the transit time between segments, according to the relation

$$t = L/\mu E$$

where t = transit time, seconds

L = distance between segment mid-points, cm

E = applied field, volts/cm

μ = mobility of holes, $\text{cm}^2/\text{volt-sec.}$

Data for a bridge of n -type germanium of resistivity about 20 ohm-cm are given in Table I. The transit distance, L , after a small correction

TABLE I

E volt/cm	Delay micro sec.	Bridge Temp. °K.	Mobility $\text{cm}^2/\text{volt-sec.}$	Transit Time micro sec.
10	21.1	298	1700	18.0
14	15.7	299	1690	12.9
20	11.8	301	1670	9.1
30	9.2	305	1640	6.2
40	9.3	313	1580	4.8

for reduced field across the side arm, was taken as 0.305 cm. As noted in the table, the bridge temperature rose somewhat at the higher bias values, and the assumed values of mobility have been modified according to the inverse three-halves power of the absolute temperature. The delay required for optimum correlation is shown in the second column of the table, and the calculated transit time between segments in the last column. It is seen that the two are in reasonably good agreement, especially at low fields. When the direction of the field is reversed, an equal delay is required, but in the opposite channel of the measuring circuit, as would be expected. Here, again, we have experimental evidence supporting the noisy hole injection hypothesis. The cause of the discrepancy shown in the table at higher fields is not understood. It is possible that trapping phenomena increase the transit time over that calculated from the mobility. There is some evidence for this sort of behavior in lifetime experiments, but to date there does not seem to be enough information for any estimate of magnitude of such an effect.

VI. GENERAL COMMENTS

These studies of electrical noise in semiconductors leave little doubt that the noise is closely related to the behavior of the minority carriers.

It is not yet clear whether the noise is a surface or a volume property of the material, but it is well established that the surface properties have an important connection with the magnitude of the noise. From some of the experimental work it seems likely that the generation and recombination processes are separate and have different noise properties. Because of the nonequilibrium situation, this does not violate the principle of detailed balancing. It seems probable that a more complete understanding of the generation and recombination processes and a clearer picture of the origin of noise in semiconductors may be expected to develop together.

VII. ACKNOWLEDGEMENT

The analysis leading to the theoretical relations between noise and magnetic field is the work of W. Shockley and H. Suhl, under whose direction the calculations leading to the curves of Figs. 10-11 were carried out. The continued interest of Dr. Shockley in the experimental work has been invaluable. The author is indebted to many associates for helpful discussion of certain problems, and also for the construction of many of the devices and materials which entered into the experimental work.

APPENDIX

Suppose that a source of holes located at a point x_0 in a filament produces a fluctuating current of holes of rms value J_1 in a specified frequency band. The hole current is swept down the filament by a field E and is assumed to decay exponentially according to the relation

$$J = J_1 e^{-(x-x_0)/\ell} \quad (1)$$

where the life path ℓ may be expressed in terms of drift velocity v , hole mobility μ , and lifetime τ

$$\ell = v\tau = \mu E\tau.$$

Assuming that the frequency of measurement is low enough to justify neglecting the hole transit time, the noise voltage due to holes from a single source is proportional to the number of holes present in the segment. This is obtained by integrating (1) over an appropriate range

$$\begin{aligned} dv &= J_1 \int_a^b e^{-(x-x_0)/\ell} dx \\ &= \begin{cases} K_1 e^{x_0/\ell} [e^{-a/\ell} - e^{-b/\ell}] & x_0 < a \\ K_1 [1 - e^{-(b-x_0)/\ell}] & a < x_0 < b \end{cases} \quad (2) \end{aligned}$$

where K_1 is an omnibus constant which will cancel out in the final result.

Under the assumption that the sources are statistically independent, the total voltage squared is obtained by integrating the square of (2) over all the sources.

$$\begin{aligned}\overline{v_1^2} &= \overline{v_2^2} = K_1 \int_{-\infty}^a e^{2x_0/\ell} [e^{-a/\ell} - e^{-(a+L)/\ell}]^2 dx_0 \\ &\quad + K_1 \int_a^{a+L} [1 - e^{-(a+L-x_0)/\ell}]^2 dx_0 \\ &= K_2 \left[1 - \frac{\ell}{L} + \frac{\ell}{L} e^{-L/\ell} \right].\end{aligned}$$

Similarly, the cross product of voltages in two segments, extending say from 0 to L and L to $2L$, is

$$\begin{aligned}\overline{v_1 v_2} &= K_1 \int_{-\infty}^0 e^{2x_0/\ell} [1 - e^{-L/\ell}] [e^{-L/\ell} - e^{-2L/\ell}] dx_0 \\ &\quad + K_1 \int_0^L e^{x_0/\ell} [1 - e^{-(L-x_0)/\ell}] [e^{-L/\ell} - e^{-2L/\ell}] dx_0 \\ &= K_2 \frac{\ell}{2L} [1 - e^{-L/\ell}]^2.\end{aligned}$$

From the definition of the correlation coefficient

$$\rho_{12} = \overline{v_1 v_2} / (\overline{v_1^2} \times \overline{v_2^2})^{1/2} = \frac{\ell}{2L} \frac{(1 - e^{-L/\ell})^2}{1 - \frac{\ell}{L} (1 - e^{-L/\ell})}$$

which is the desired relation, from which the solid curves of Figs. 14–16 were calculated.

REFERENCES

1. P. H. Miller, *Proc. Inst. Radio Engrs.*, **35**, pp. 252–256 (1947).
2. J. A. Becker and J. N. Shive, *Elec. Eng.*, **68**, pp. 215–221 (1949).
3. R. M. Ryder and R. J. Kircher, *Bell System Tech. J.*, **28**, pp. 367–400 (1949).
4. R. L. Wallace, Jr., and W. J. Pietenpol, *Bell System Tech. J.*, **30**, pp. 530–563 (1951).
5. W. Shottky, *Phys. Rev.*, **28**, pp. 74–103 (1926).
6. M. Surdin, *Journal de Phys. et le Rad.*, **10**, 188–9 (1939) do **12**, pp. 777–783 (1951).
7. G. G. Macfarlane, *Proc. Phys. Soc.*, **59**, Pt. 3, 366–374 (1947) do **B63**, pp. 807–814 (1950).
8. J. M. Richardson, *Bell System Tech. J.*, **29**, pp. 117–141 (1950).
9. F. K. duPré, *Phys. Rev.*, **78**, p. 615 (1950).
10. A. Van der Ziel, *Physica*, **16**, pp. 359–372 (1950).
11. W. Shockley, G. L. Pearson, J. R. Haynes, *Bell System Tech. J.*, **28**, pp. 344–366 (1949).
12. W. L. Bond, *Phys. Rev.*, **78**, p. 646 (1950).
13. G. B. Herzog and A. Van der Ziel, *Phys. Rev.*, **84**, pp. 1249–1250 (1951).
14. W. Shockley, *Electrons and Holes in Semiconductors*, Van Nostrand (1950).

Important Design Factors Influencing Reliability of Relays

By J. R. FRY

(Manuscript received June 13, 1952)

Relays are produced by a large number of manufacturers in this country. When we survey their product, we find that there are many kinds and varieties. They differ widely as to their size, shapes and configurations. Many of these differences are dictated by the requirements of the task they must perform and by the environments under which they must work. Other differences are brought about from considerations of cost and by the design and fabrication techniques the particular manufacturer employs. However, all relays have a common objective. For whatever use they are employed, it is highly desirable that they be reliable. They are expected to function each time they are called upon without failure and over the expected life of the equipments in which they are used. This paper deals with the more important design factors which all relays have in common that greatly influence their reliability of performance. Contact spring pile-up stability and the importance of strength of screws, insulating materials with low cold flow and moisture absorption, and manufacturing procedures and controls to achieve this end are discussed. Coil construction so as to minimize the occurrence of open windings due to corrosion of the wire and breakage of the lead-out wires is dwelt upon. Contact reliability and how it is affected by the material used, its size and shape, the method of actuation, the presence of contaminating vapors, and single versus twin contacts are discussed. The degree by which magnetic materials change their magnetic properties with age and treatments for alleviating this effect are described. The importance of adequate structural design so that the relay will be rugged and remain stable so that its performance is substantially unaffected by wear, shock and vibration is stressed. Methods of test to determine how well the relay meets these objectives are described.

Although a relay is conceptually a simple device, the wide range of conditions under which relays are required to operate, the many different characteristics they must have, and the complete dependence placed

upon them in many circuit applications, make them a subject of continuous study.

In the telephone industry, for example, the completion of a single call may bring into play a thousand or more relays. While their principal function is to close electrical contacts, there are many facets to the problem of doing this satisfactorily. Relays are produced by many manufacturers in this country. When we survey their product we find that there are many kinds and varieties. Shapes, sizes and configurations of relays may differ in accordance with the requirements of the tasks they must perform, and the environments under which they may work; other differences may be brought about by cost considerations and by design and fabrication techniques of the manufacturer.

All relays, however they may be used, have one common objective — they must be reliable. They are expected to function each time they are called upon, should do this without failure, and should continue to do so over the expected life of the equipment in which they are used. It is the purpose of this paper to discuss the more important factors that are common to all relays and which have considerable influence on their reliability of performance. The design considerations discussed in this paper are presented in the following order.

- (1) Contact Pile-up Stability,
- (2) Coil Construction,
- (3) Contact Reliability,
- (4) Magnetic Stability, and
- (5) Structural Stability.

CONTACT SPRING PILEUP STABILITY

Stability of the contact spring pile-up contributes in a large degree to the reliability in performance of the relay. Since contact springs may be assembled into pile-ups of from two springs to a dozen or more, they must be secured so that they will not shift position during the life of the relay, even when it is subjected to relatively large changes in temperature and humidity, and to vibration and shock during shipment, installation, wiring, and under operating conditions. It is also important that the dimensional relations between the contact ends of the springs and the actuating members of the relay do not change appreciably; otherwise, changes in contact separation, contact follow, contact pressure, and operating and releasing current values may cause faulty operation of the relay.

Insulators for securing the springs should be made of materials having low cold flow and moisture absorption characteristics; in telephone relays,

the better grades of phenol fibre have been found adequate. They should have generous clamping surfaces, so that when the spring pile-up is clamped under force, high pressures on the insulators are avoided — thus minimizing cold flow and keeping well below the crushing strength of the material.

Pile-up screws, clamping plates, and screw threads should be proportioned such that permanent deformation under any condition does not take place, i.e., the maximum stress does not exceed the elastic limit of the metal. It has been found advantageous to use high tensile strength steel for these parts — a tensile strength of 100,000 lbs per square inch or higher.

In the manufacture of relays, the desired pile-up tightness requires certain procedures and controls. Insulators are baked in an oven at a temperature of about 150°F for a minimum of 24 hours and assembled in the relay while in the dry condition. During assembly, prior to tightening the screws, the pile-up is pressure clamped under a hydraulic or air powered fixture to a controlled force of 1,300 lbs to 3,200 lbs, depending upon the size of the relay; while under this pressure, the screws are tightened using a controlled torque. To assure that the processes and materials are under control, the relay is then tested on a “no-go” basis in a fixture that applies a definite force on the contact springs in a direction to rotate them in the pile-up.

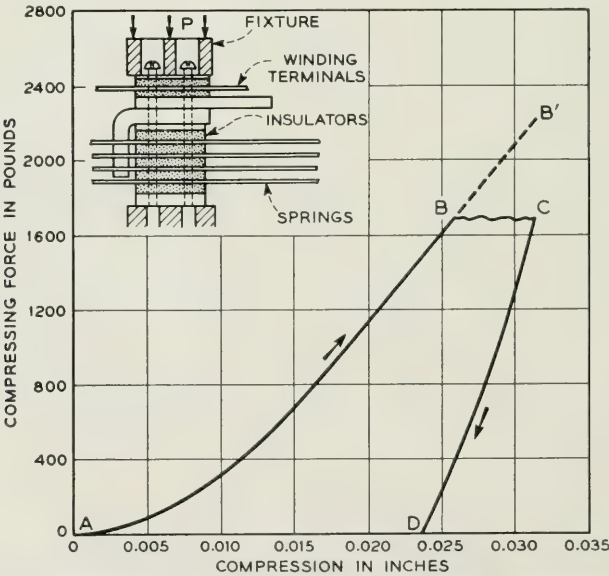


Fig. 1—Compression cycle of a relay contact spring pile-up assembly.

Pressure clamping during manufacture enables the spring pile-up to better maintain its adjustment through cycles of humidity and drying, and to prevent displacement during installation and wiring. The action of a pile-up under this compression is illustrated in Fig. 1. Curve AB represents the application of a 1,700-lb force to the pile-up by a power driven fixture prior to tightening of the pile-up screws. At the start, the relationship is not linear due to "nesting" of the parts, but a linear slope is soon reached, representing the stiffness of the pile-up without the screws. At the point B the two screws are tightened with a controlled torque; further compression takes place, indicated by the jagged line BC. An estimate of the tension put into the screws by this tightening operation can be made by extrapolating the curve AB to the point B^1 , vertically above the point C. When the pressure fixture is released, the pile-up tends to expand and follows the line CD. The slope of this line represents the combined stiffness of pile-up parts and screws, which makes it stiffer than the original compression slope. When the pile-up is released, an equilibrium point is reached where the tension in the screws equals the force with which the pile-up tends to expand.

A series of measurements on a typical relay pile-up screw and clamp plate assembly is shown in Fig. 2 to illustrate the stress-strain relationship when a force is applied axially to put the screws under tension. As force is applied, Hooke's law is followed up to the point A; strain is proportional to the stress and no permanent deformation takes place. Beyond point A the elastic limit of the metal is exceeded and permanent deformation begins. When a point B is reached, somewhere below the breaking point of the screw, and the force is released, a permanent deformation results. Note that, in Fig. 2, the high strength screw will permit a higher screw tension without deformation – and its resultant looseness of the pile-up – than will the lower strength screw.

Analytical methods are available for estimating the range of screw tensions that exist during the life of the relay. By taking into account the known cold flow characteristics of the insulators with the relay in the dry state, a minimum value can be estimated. It should be of sufficient value to hold the springs securely in place. By considering the conditions that obtain in the humid state, maximum value of tension can be predicted, which should not exceed the strength characteristics of the materials used.

To determine how well the design objectives for stability are being realized, accelerated laboratory tests are made upon the relay, and from these results predictions can be made as to its performance during its life. In the telephone system, relays are generally subjected to repeated

cycles of alternate humid and dry environments over the years. Humid conditions exist during the summer season, followed by a dry exposure during the winter months when the central offices are heated. Experience has shown that a relay exposed for six days to 90 per cent relative humidity at 85°F will be comparable to those observed in service in humid localities. Although the six day exposure is admittedly an accelerated test, the dimensional changes produced are approximately the same as those caused by the accumulating effect of fluctuating humidity during the entire season. Similarly, a period of six days exposure to 120°F produces the same effect of drying as that which occurs during the heating season.

By making careful measurements on an adjusted relay of such important parameters as operate current, releasing current, contact separation, contact spring tension, armature back tension, stud gaps, etc., and then subjecting the relay to repeated cycles of humid and dry conditions, repeating the measurements after each exposure and noting the changes, a good appraisal of the relay can be made. A repetition of the test will reproduce the same pattern of results as the first cycle unless permanent deformation of the materials in the relay has taken place.

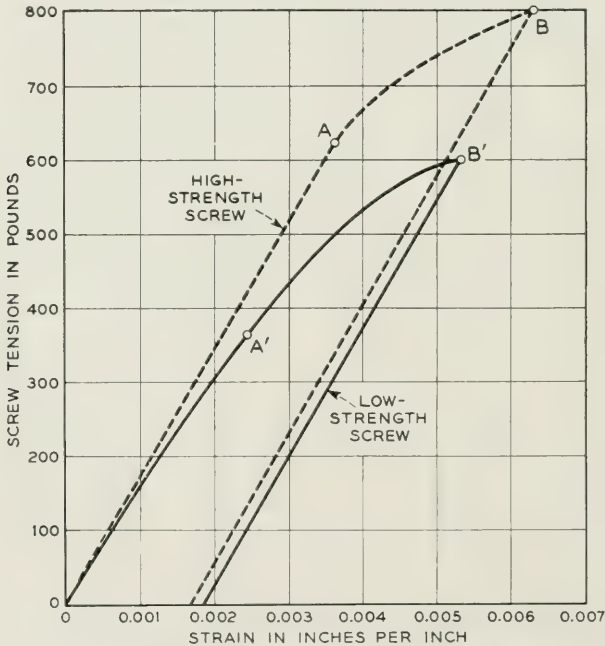


Fig. 2—Strength of relay pile-up screws.

COIL CONSTRUCTION

If, for any reason, the winding of a relay should become open during its life, usefulness of the relay ceases. One of the most prevalent causes for failure of this kind can be due to so-called corrosion of the wire. This is not corrosion in the ordinary sense of the word, but is caused by an electrolytic action within the coil when an electrical potential is applied to the winding. This action can take place only in the presence of moisture. If there are any active impurities in the insulating materials intimately associated with the copper wire, an electrolyte is formed and disintegration of the wire proceeds to the point where failure may occur. This trouble is accentuated with coils using small diameter wires, because with the smaller cross-section of copper, failure of the section will occur in a shorter period of time.

There are two methods of approach to minimize corrosion failure. One is to thoroughly dry the coils, and in this condition seal them in a potting compound, which prevents the entrance of any moisture into the coil, or enclose the coil in a hermetically sealed chamber. For relays this is an expensive and cumbersome way to overcome corrosion troubles. The second and more practical method is to use, in the construction of the coil, insulating materials that are chemically inert and free from corrosion promoting impurities.

For many years, studies were made with a view towards eliminating the occasional corrosion failures of fine wire windings which occurred under unfavorable atmospheric and circuit conditions. Although improvements were effected by the use of the better grades of phenol fibre for spoolheads and waxed varnished papers for core and winding insulation, an entirely satisfactory coil was not achieved until the use of cellulose acetate insulation was adopted. This material, in thin sheet form, can be applied to spool-wound coils, where the coils are wound individually, and the so-called "filled" coils, where a multiple number are wound simultaneously on automatic winding machines. The coils are wound on a mandrel, as many as twelve individual coils per mandrel, with a thin sheet of insulation between the layers of wire. The "stick" of coils is wound so as to leave a small separation between coils to provide insulation at the ends of the coils and to permit cutting the stick into individual coils, after which they are assembled to relay cores using conveyor assembly methods. This results in not only a more economical coil, but in a higher quality coil as well. The thin sheet of insulation between the layers of wire, generally not provided on the spool wound type of construction, eliminates the occurrence of short-circuited turns.

The degree of improvement obtained by the use of cellulose acetate over the materials formerly used is shown in Fig. 3.¹ This is the result of an accelerated corrosion test adopted as a means of obtaining data on proposed constructions, using as a basis for comparison, the performance, in this test, of the earlier spool wound construction. The latter had given satisfactory service in the field, only occasional failures occurring under the most severe circuit and atmospheric conditions. This test is made with double or triple wound coils since these represent the most

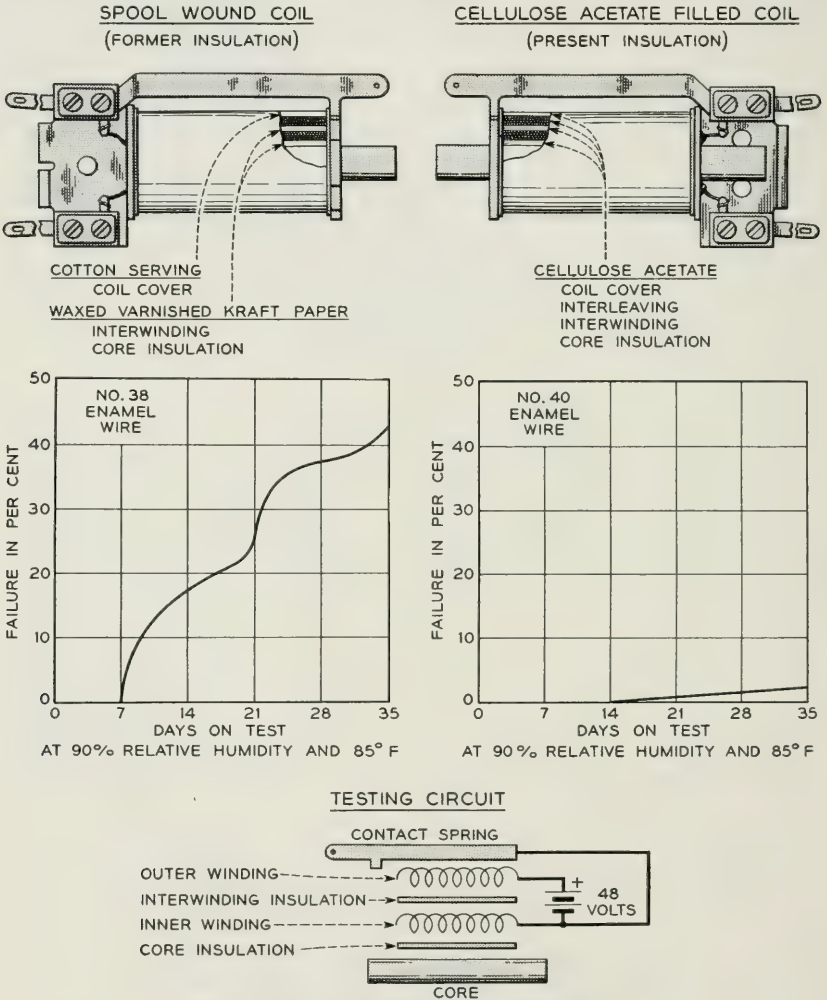


Fig. 3—Corrosion tests on relay windings.

serious conditions. A group of coils are subjected to 90 per cent relative humidity at 85°F with negative potential applied to the inner winding and positive potential to the outer winding. No current flows in the windings. Depending upon the type of apparatus in which the coil is used, other parts of the structure may be made positive or negative to simulate actual service conditions. When electrolytic action takes place, copper is always eaten away from the positive electrode. Consequently in practice where there is a choice, it is better to keep the winding negative with respect to its surroundings. During a 35 to 40 day period, continuity checks are made periodically using a Wheatstone Bridge having a battery supply of $1\frac{1}{2}$ volts in a series with 10,000 ohms. This method of test does not provide high enough voltage nor permit flow of sufficient current to establish continuity through a minute length where the wire may be corroded through, nor does it cause a reduced section of wire to burn out. In other words, this method of test does not restore continuity in a corroded through section nor does it destroy metallic continuity. Thus more consistent results are obtained than if higher voltages or currents were to be used. The marked superiority of the cellulose acetate insulated coil is apparent, and experience with its use in service has shown that corrosion failures have been eliminated.

From time to time the question arises as to how the cellulose acetate insulated coil compares with coils vacuum impregnated with a varnish and employing other types of insulation for use in equipment for the Armed Services where atmospheric conditions are more severe than those ordinarily encountered in the telephone plant. Frequently specifications for these applications require impregnation of the windings. Tests have shown that impregnation will extend the life of a coil employing inferior materials, but that corrosion will take place in a shorter period than where cellulose acetate is used throughout without impregnation.

Results of such tests are shown in Fig. 4 where the various groups of coils were kept in a humidity chamber at 95 per cent relative humidity. The temperature within the chamber was raised and lowered between limits of 85° and 150°F in cycles so as to produce severe condensation on the coils. Each cycle, plotted as abscissae represents 24 hours of exposure. The top two curves IIIA and IIIB show the failure rates of two groups of coils constructed exactly alike except that one group was impregnated while the other was not. They were cellulose acetate filled coils, but used lead-out wires insulated with commercial grades of braided cotton. The two curves IIA and IIB represent the results on two groups of spool wound coils having cellulose acetate core and interwinding insulation, but provided with vincellatate muslin covers and red-rope paper

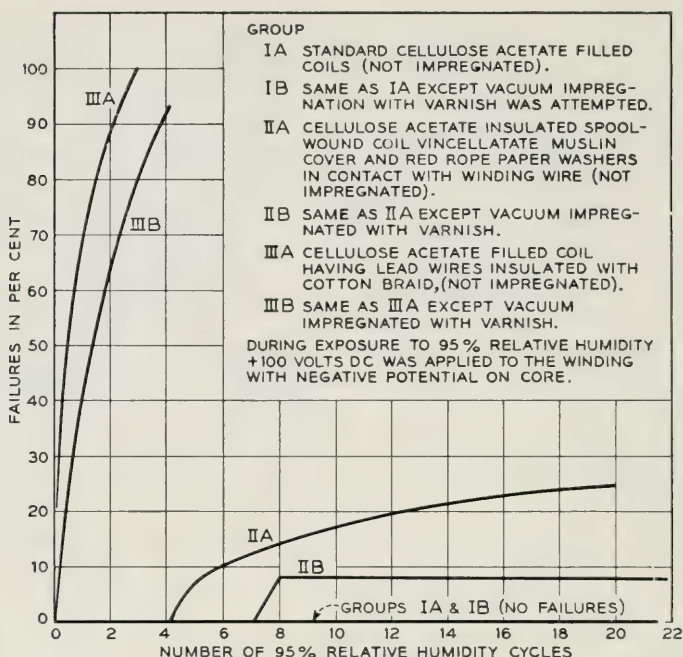


Fig. 4—Corrosion comparison of impregnated and non-impregnated relay coils.

winding washers. Likewise, one group was impregnated while the other was not. Fifth and sixth groups of coils having cellulose acetate insulation throughout with and without impregnation were exposed and there were no failures at the end of the test. This shows that the corrosive effects of impure materials can be retarded, but not overcome by resorting to impregnation. In general, impregnation of relay coils is not desirable because of the risk of contaminating the vital working surfaces of the relay.

In the normal operation of a relay when the circuit through its winding is opened, a transient voltage, which may reach hundreds of volts is generated across the winding terminals by the collapsing flux. If the insulation between the lead-out wires or between the lead-out wires and the end turns of the winding is not adequate, electrical breakdown causes arcing and repeated operation of the relay may cause ultimate disintegration of the wire and consequent failure of the relay. It is important, therefore, to design the coil so that lead-out wires under all conditions are properly spaced, and to provide adequate insulation between those portions of the winding where high voltages can exist. A test has been de-

vised for use during manufacture which will detect an incipient failure of this kind. By pulsing the relay in its normal fashion the self-generated coil voltage on breaking the circuit can be observed on a cathode ray tube; any deviation in this voltage caused by momentary breakdown or shorted turns can be detected readily.

Another source of coil failure is lead breakage, caused principally by fatigue of the small copper wires. Copper has a low fatigue strength and if it is subjected to repeated bending strains, eventually it will break. As the relay operates and releases, impact of the armature against the core and backstop causes shock and vibration of the coil; coil construction therefore needs to be such that strains are not imposed upon the fine wire by this motion.

On spool wound coils, being individually wound, the fine winding wires can be reinforced by stranded lead-out wires for connecting to the relay terminals. Besides, the coil is generally wound tightly on the relay core. These factors, to a large extent, preclude lead breakage. With the cellulose acetate filled coil it is desirable to bring the winding wire directly to terminals on the terminal spoolhead to which the end of the coil is later bonded. Since the filled coil must slide loosely over the core for assembly reasons, it can have a small lateral motion at the non-terminal end. To eliminate this movement a "motion limiting" washer has been provided to fit snugly over the core and which is bonded to this end of the coil.² The washer and the way it is used is illustrated in Fig. 5. In assembly, the thin cellulose acetate faced phenol fibre washer and the non-terminal spoolhead are forced over the knurl on the core and the washer is bonded to the end of the coil. The tight fit of the washer on the knurl is the feature that prevents lateral movement of the coil. There is always some slight shrinkage of the cellulose acetate filled coil in the

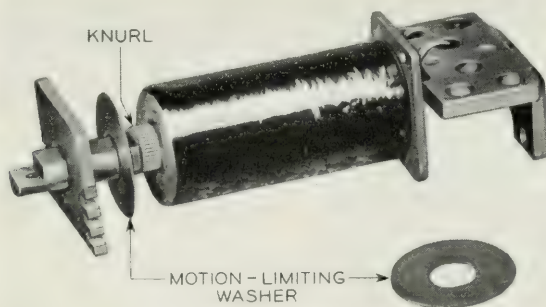


Fig. 5—Relay coil employing a motion limiting washer to prevent lead breakage.

longitudinal direction, but the washer can move with the coil, although eliminating lateral movement. Use of this washer has practically eliminated fatigue lead breakage.

CONTACT RELIABILITY

Since the opening and closing of contacts are the prime objectives of a relay, it is extremely important that the contacts themselves are made reliable. To realize these objectives, several factors should be taken into consideration. First is the contact material. While much could be said regarding the behavior of contact metals, space does not permit more than a brief treatment. The contact should maintain reasonably low resistance and under the environment in which it is used, be able to withstand the erosion. Electrical resistivity of most metals is low enough to be satisfactory from the resistance standpoint, but unfortunately, most of them develop tarnish or corrosion films when exposed to the atmosphere, thus increasing the contact resistance and rendering them unsuitable for a contact. These metals are sometimes referred to as "base" metals, and include aluminum, brass, bronze, copper, chromium, nickel and stainless steel. There is a much smaller group of metals known as "noble" or precious metals, such as platinum, palladium, gold and iridium. These are relatively free from the tendency to tarnish and will maintain low contact resistance. Alloys of these metals and certain alloys in which silver is included are widely used in the telephone plant. Pure silver is also used and is attractive because of its low cost; however, it has a tendency to form high resistance tarnish films and therefore has limitations in its use. It is employed in signaling circuits where the contact makes or breaks current. Its contact resistance remains low because the films that form on the silver are broken down or destroyed by the arc. It is not employed in circuits carrying voice currents on account of its tendency to introduce noise.

Enough metal must be provided to give satisfactory life. Each time a contact makes and breaks an electrical circuit, a small part of the metal may be lost, so that life may be considered roughly proportional to the volume of metal available for erosion. The pair of contacts must have sufficient height to provide enough contact spring clearance to allow for spring adjustment and to insure that the springs will not touch during the normal operation of the relay. At least one contact of a pair must be large enough, that is, present a sufficiently large target area, to insure full registration of the contacts with normal manufacturing variations of the position of the contacts on the springs and with the variations in alignment of the springs during assembly.³

In the early days of relay design, contacts were attached to the springs by riveting. In fact, many of the relays manufactured abroad today are made in this manner. In this country, during the past 30 years or more, spot welding has largely replaced the riveted construction. This was done for economy reasons. Spot welded contacts, unless carefully controlled during manufacture, may not be so reliably attached as the riveted contacts and the likelihood of the contacts dropping off during the life of the relay may be greater. It has been found in welding the millions of contacts required in the telephone system that close control is required in several factors affecting the quality of welds obtained with any given material. Important factors are cleanliness of the welding surfaces, pressure between electrodes, welding current and the time during which the current is applied. In order to insure that these factors are at all times under control, and since the consequences are rather grave when poor quality welds are produced, it has been found desirable to institute frequent inspections of the quality of welds at each welding machine on a sampling basis. Periodically a small number of contacts are subjected to a destructive test in which the force required to shear off the contact is measured. In this manner, any deterioration in the quality of welds can be detected early, and corrective measures can be applied.

One type of failure sometimes experienced with relays is contact locking.⁴ When a contact is closed by the operation of the relay it may become mechanically locked to the contact member with which it is engaged and fail to open when the relay is released. As a result of arcing as the contact closes and opens an electrical circuit there is a transfer of metal from one contact to the other. This building up and wearing away leaves both contacts roughened. If the opening and closing motion were along a perpendicular to the face of the contacts, this roughening would ordinarily have little effect. But with a slight sliding or rocking motion at the contacts after they come into engagement, small projections on one contact may lock mechanically in a cavity on the other and thus prevent the contacts from opening when they should. When contacts have locked, measurements have shown that forces in excess of 100 grams may be required to separate them.

This kind of failure can be avoided by employing an improved method of spring actuation.⁵ This is illustrated in Fig. 6. At the top of the figure is a stud actuated contact spring assembly. The spring carrying the moving contact is tensioned away from the fixed contact member and exerts a force to hold the armature against the backstop when the relay is unoperated. A stud, moved by the armature, presses against the moving contact spring a short distance back of the contact to close the contact

when the relay is operated. As will be noticed, the further deflection of the contact spring necessary to obtain the required contact force after closure, causes the moving contact to slide and rock slightly on the fixed contact. For the reasons previously mentioned such a contact is prone to lock when the conditions are favorable.

Now, the bottom of the figure shows a card actuated contact spring arrangement. Here a phenol fibre card is employed to operate the contact instead of a stud. The moving contact spring itself is pretensioned against the fixed contact to give the desired contact force. The card is held by two card springs that are tensioned away from the fixed contact in a slightly greater amount than the moving contact spring is tensioned toward it. As a result, when the relay is unoperated, the card holds the make contact away from the fixed contact. When the relay operates, the armature pressing against the top of the card pushes it toward the fixed contact and allows the contact to close. It is quite apparent that with this actuation the moving contact engages the fixed contact without

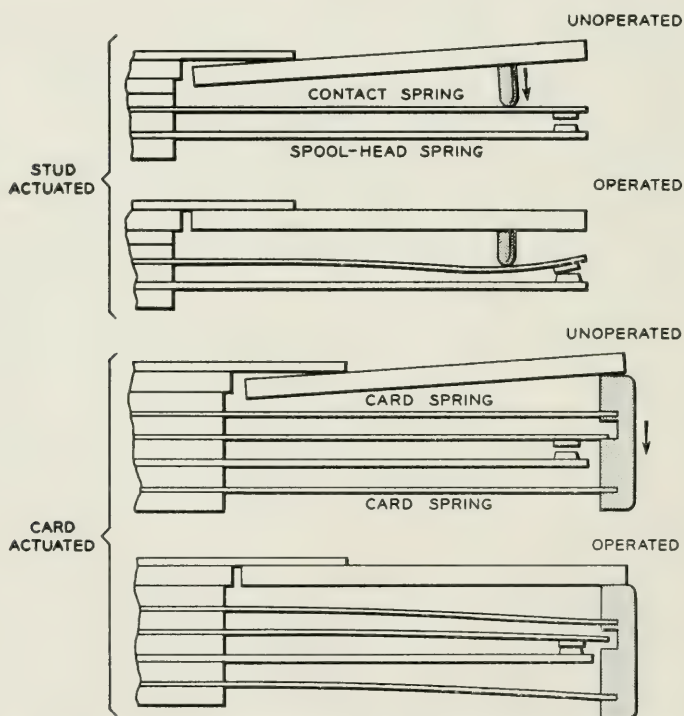


Fig. 6—Two methods of contact spring actuation and their influence on contact locking.

sliding or rocking on it and the tendency to lock is thus largely avoided. The likelihood of locking is further decreased because the restoring force is greater and is applied closer to the contact, and because of the impact of the card on the spring when the relay releases. With the contact closed there is clearance between the contact spring and the bottom of the slot in the card, and thus when the card hits the spring in opening, it is already moving and has acquired kinetic energy. This energy is available on impact to overcome any locks which may have occurred.

Recent studies have shown that erosion of electrical contacts on closure is due almost entirely to an arc occurring, in most cases, before the contacts touch.⁶ When there is no arc there is no erosion. It has also been observed that the occurrence of an arc between the approaching contacts is influenced by operation in the presence of various organic vapors; for example, benzene derivatives.⁷ The effect of such operation is to permit arcing on lower currents than is the case with clean contacts and results in increased erosion rates. This is true for noble metal contacts, and when so exposed they are said to have become "activated". A metal surface which has been activated by organic vapor remains active indefinitely if there is no arcing at the surfaces. With continued operation and accompanying arcing, the activating material is burned away, and the surface returns to the inactive condition, provided no contaminating vapor is present.

Some materials used in relays may give off organic vapors which can aggravate the arcing at the contacts. A series of experiments has been made by placing various materials under test in a small glass enclosure and proceeding to find if, and at what elevated temperature, vapor from the materials will give rise to arcing on "make", with contacts that are operating within the enclosure. The materials tested varied widely in their effects upon arcing at the relay contacts. In the solid organic group, they ranged from polystyrene, which produced arcing at room temperature, to teflon, which did not cause arcing until heated above 200°C.

The precise correlation between the results of these tests and the changes in erosion rates, which occur when these materials are used in the relay construction, has not yet been established. However, they may be used as an aid in the choice of such materials. Cases have come to our attention both in the laboratory and the field where the erosion rates of relay contacts operating in confined chambers were many fold those which occurred when the relays were operated in the open. This was at least partially ascribed to the presence of contaminating materials known to be present.

Another type of failure that is quite generally experienced in relay

operation is "open" contacts due to small insulating particles present in the atmosphere becoming trapped between the contacts. This causes high resistance or open circuit and consequent circuit failure. Many attempts have been and are being made to reduce "open" contact troubles. Examples are filtering the air supply to the central office, enclosing the relay equipments in closed cabinets, pressurizing the enclosing cabinets, covering smaller groups of contacts by independent covers, employing twin contacts rather than single contacts, and enclosing the relay or its contacts in a hermetically sealed chamber. Even going to the extreme of completely isolating the relay from its surroundings is not a complete answer. There is always the possibility of failure by wear particles generated within the enclosure by the relay actuation.

The most widely employed method to reduce dust failures in the telephone plant is the use of twin contacts in combination with some of the above mentioned types of enclosures. If the incidence of dust failures followed the laws of probability, then elementary considerations would lead us to predict that if single contacts failed at the rate of once in 1,000 operations, the simultaneous failure of the two such contacts comprising the twin would be once in 1,000,000 operations. This is the so-called "square" law. However there are a number of reasons why this is not realized in practice and why the figure of merit for the twin contact is very much less than that indicated by the "square" law. In the first place, when foreign matter becomes lodged on a contact, it seldom falls out on the first subsequent operation, but will require a number of operations before it cleans itself; in fact, it may remain inoperative indefinitely. When this happens to one of the twin contacts, during this period of time, the twin contact is no better than the single contact. In practice, twin contacts are generally used with the same total force as the single contact, being nominally divided equally between the two contacts. This reduction in force per contact on the twin contacts is of considerable importance in reducing its effectiveness. Relay designs employing twin contacts that have been used in the past do not have complete mechanical independence of the two members to which the contacts are attached. Foreign material or protrusions under one contact can adversely influence the performance of its mate. In a new design of relay which is about to go into production, the design criterion that twin contacts to be most effective should be completely and mutually independent has been met. Laboratory tests and field experience obtained to date show a marked improvement over the former designs in regard to the incidence of open contact failures.

Tests have been made repeatedly in the laboratory for comparing the

performance of twin contacts with single contacts, and arriving at a figure of merit. Data have also been collected from relays in service in the telephone plant on the basis of numbers of found open troubles on both types of contacts. As might be expected the results varied widely, with the twin contact being superior by a factor of anywhere from 3 to 100 with perhaps 10 as a reasonable figure.

MAGNETIC STABILITY

Magnetic materials in relays have been found to change in their magnetic characteristics with time and temperatures to which they are subjected in their normal usage. This effect is known as aging. The direction of the change is such as to decrease the permeability and increase the coercive force of the material. The degree of change in certain applications, such as relays in marginal and time delay circuits, may be so large as to be of serious concern.

A high grade of magnetic iron which has been extensively used in the telephone system has been found to age considerably under conditions simulating operation in the plant. Aging of iron is attributed to the precipitation of impurities such as carbon, nitrogen, and oxygen. The solubility of these elements decreases with decreasing temperature. When iron is cooled from a high temperature, impurities, such as carbides and nitrides, do not have sufficient time to precipitate completely, so a super-saturated solid solution is produced. Consequently the impurities tend to continue to precipitate slowly at low temperatures where the diffusion rate is extremely slow, and internal strains are produced which affect the magnetic properties.⁸

It has been found that if these parts are annealed in atmospheres of dry hydrogen instead of the ordinary "pot" anneal, this aging effect is greatly reduced. Not only is the aging effect reduced to where it is of no great engineering importance, but the magnetic properties of the material are improved. The maximum permeability is increased and the coercive force is decreased both by a factor of about two. The use of relays in critical applications is thus greatly enhanced.

The degree by which magnetic materials change by aging may be determined readily by laboratory tests. Long time aging effects can be simulated by baking ring samples of the material or the relays at 100°C for several hundred hours and measuring the magnetic properties of the ring specimens or the operating characteristics of the relays before and after aging. For "pot" annealed magnetic iron the effect of such aging is to decrease the maximum permeability by about 50 per cent and to approximately double the coercive force. When the iron is hydrogen

annealed, the corresponding changes caused by aging will be about a 15 per cent decrease in maximum permeability and 15 to 20 per cent increase in coercive force.

The improvement in aging effect on relay performance obtained by the hydrogen treatment is illustrated in Fig. 7.² This was obtained on a design of relay having a closely coupled magnetic circuit for use in time delay circuits. The ordinates show the change in residual grams; hours of aging are plotted as abscissae. Residual grams represent the force with which the armature is held attracted to the core by the residual flux remaining in the magnetic circuit after the electrical circuit through its winding is opened. This force is a measure of the coercive force of the magnetic material. As was noted, the effect of aging is to increase the coercive force and hence the residual grams. For this relay, a change in residual grams will cause a change in the delay time of the relay under a given adjustment and is therefore important.

How hydrogen annealing improves the pull characteristics of a relay is shown in Fig. 8. This was taken on a relay designed for sensitive and marginal circuit applications. The curves show the grams pull, plotted as ordinates, produced on the relay armature at a given air gap by various values of ampere turns on the relay plotted as abscissae. The ability of the hydrogen treated relay to operate given loads on considerably smaller currents is obvious. This improvement is due to the higher permeabilities obtained by the hydrogen anneal.

There are other magnetic materials available for use in relays and in which the aging effect is practically non-existent or is considerably smaller than that just described. Several kinds of nickel-iron alloys known as permalloy are widely used in the telephone system where their

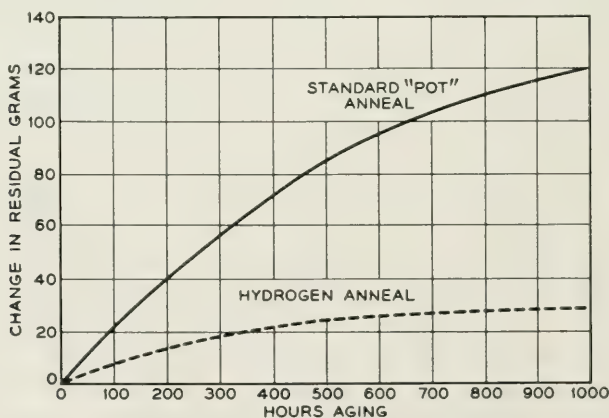


Fig. 7—Improvement in aging effect by hydrogen anneal.

excellent magnetic properties are needed in difficult applications. They are substantially non-aging. Low silicon-iron alloys are being more widely employed. They have good magnetic properties and the aging effect is small. Where the intrinsically inferior magnetic properties of low carbon steel alloys, such as SAE 1010, can be tolerated they are used. While initially they have poorer magnetic characteristics than magnetic iron, their aging effect is considerably smaller.

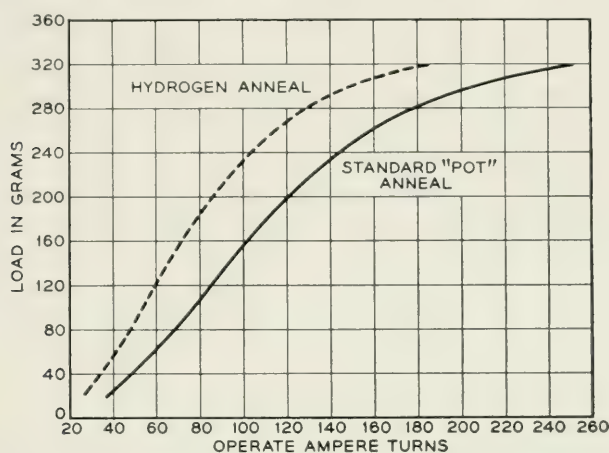


Fig. 8—Improvement in relay performance by hydrogen anneal.

STRUCTURAL STABILITY

Response of a relay depends upon the value of the magnetic force of attraction produced between the armature and core when the coil is energized, and upon the magnitude of the mechanical forces acting upon the armature. To keep the response constant during the life of the relay, it is essential that the relationships between these two forces be not changed. The force of attraction varies approximately inversely with the square of the length of the air-gap between the armature and core. Since this distance is usually small, any small change will have a relatively large effect on the pull. The core and armature, together with their associated members, should be of stable design and secured in such fashion that their dimensional relationships remain unchanged when the relay is subjected to shock, vibrations, and stresses incident to attaching the relay to its mounting. The design of the structure and the thermal coefficients of expansion of the materials used should be such that deformation does not take place when the temperature is varied throughout the operating range.

Moving parts, such as the armature and its suspension, together with the associated actuating members and springs, should move freely under all conditions without binding or friction. Since friction is inherently a variable quantity and difficult to control, it should be kept as small as possible, otherwise it will be a cause for instability. If the friction component is an appreciable part of the total load, the relay will be unsatisfactory, particularly for marginal operation. The friction part of the load on the moving system of a relay can be determined readily by an improved measuring technique which automatically plots the force required to move the armature, and its displacement, as it moves from its unoperated to its operated position. This is illustrated in Fig. 9. The top curve is the force required to move the armature and operate its associated contact springs as it moves from its back-stop to its fully operated position. The lower curve is the force acting on the armature that allows it to restore to its unoperated position. The vertical displacement between the two curves represents double the friction. Since friction always opposes motion, it adds to the force required to operate the relay and detracts from the force releasing the relay. If the ideal of no friction were obtained the two curves would coincide. When

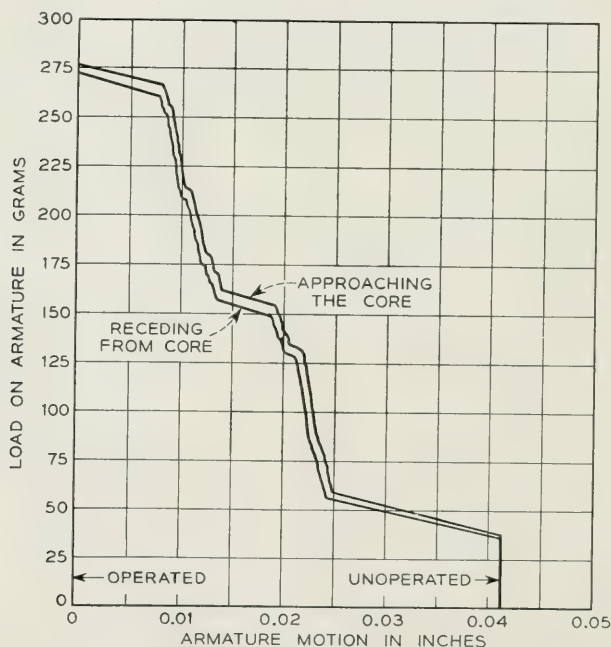


Fig. 9—Mechanical load of a relay and its friction component.

the displacement and hence the friction is large, aside from the fact that it is indicative of a rapid rate of wear, the relay would be unstable.

While the wear of the relay parts can be minimized by good design, it cannot be eliminated entirely, especially for relays required to operate a very large number of times during their life. For telephone relays the design objective is for a 40-year life. The effects of wear on performance to a great extent can oftentimes be counteracted by ingenious design. Fig. 10 is an illustration of such a case.⁵ The diagram on the left shows a moving system of a relay in which the contact springs are stud actuated. The moving springs are tensioned toward the armature and exert a force tending to open the contacts. When the armature operates, the stud presses the moving springs into engagement with the stationary springs. There is no contact force when engagement is first made and further flexing of the spring is necessary to build up the contact force to the desired value when the armature reaches its fully operated position. As the contacts and studs wear, it is apparent that the contact force and consequently the load on the armature decreases rapidly. The stud wear becomes cumulative in its effect on the outside pair of springs as more springs are added to the pile-up.

The diagram to the right shows a moving system of a relay using what is called "lift-off" card actuation. The moving springs are ten-

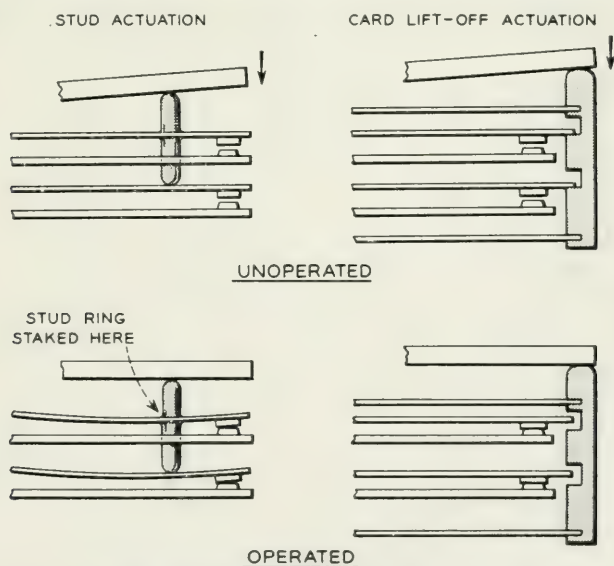


Fig. 10—Two moving systems of relays in relation to the effects of wear on their performance.

sioned, before assembly, toward the stationary spring by an amount necessary to give the desired contact force. Two supplementary springs are provided to support the card and tensioned to restore the armature and contact springs to their unoperated position. Upon operation, the motion of the card permits the contacts to close, and when engagement of the contacts occurs, the contact force reaches its predetermined value very rapidly. Further motion of the card, provided for by the width of the slot in the card, allows for wear of the contact and card without appreciably affecting the contact force or the load on the armature.

The effect of this wear on the contact force is shown in Fig. 11 for both types of actuation. For the stud actuated relay, as the contact and stud wear continues, the contact force decreases very rapidly. After 0.010 inch wear only about 6 grams remains out of an original 26 grams. This is accounted for by the fact that the combined stiffness of the moving spring in engagement with the stationary spring is 2 grams per 0.001 inch deflection. This requires 0.013 inch contact follow to establish a contact force of 26 grams when the relay is adjusted initially. For the card "lift-off" actuated relay where the moving spring had been pre-tensioned to give a contact force of 25 grams initially, after 0.010 inch wear of the contacts, the contact force will have decreased about 1 gram. This is because the stiffness of the moving spring is about 0.1 gram per 0.001 inch deflection. Card wear does not affect the contact force so long as it is provided for by the width of the slot in the card.

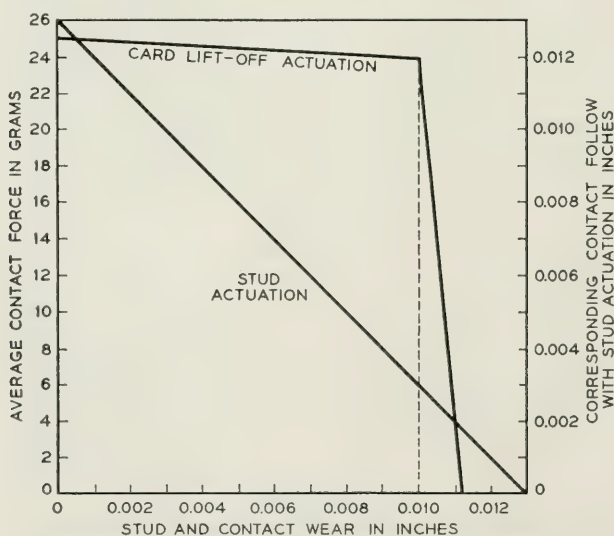


Fig. 11—Comparison of effects of wear on contact pressure of a relay.

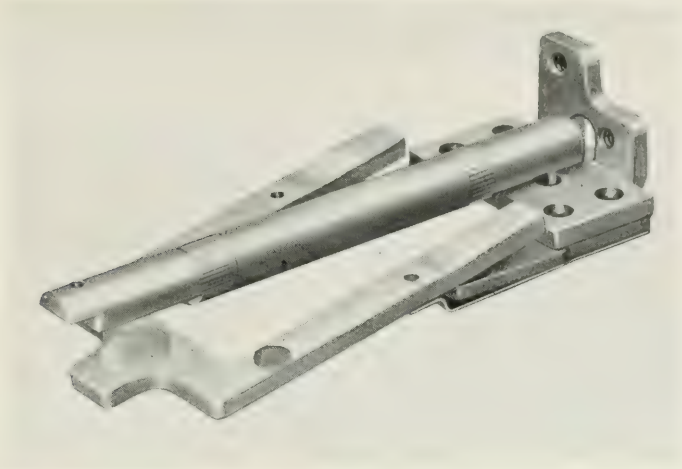


Fig. 12—Magnetic circuit of a relay having embossed pole faces.

Another instance where the effects of mechanical variations upon its performance have been largely nullified by design, is in the design of slow release copper sleeve relays. To make most effective use of the copper sleeve, which causes the delayed action, it is desirable to provide as low a reluctance as possible of the magnetic circuit when the relay is in the operated position. Instead of providing small non-magnetic separators in the air-gap between armature and the core as is usually

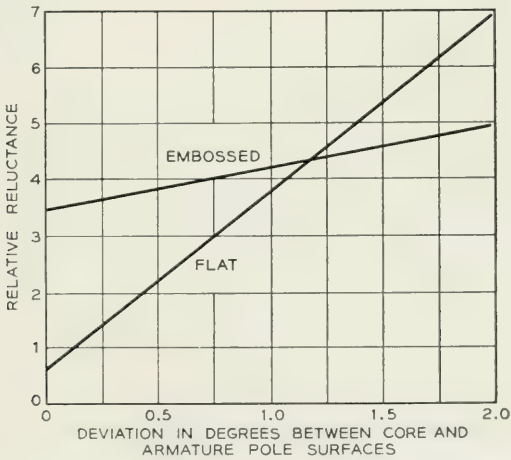


Fig. 13—Comparison of flat and embossed pole surfaces and their magnetic closed circuit reluctance with misalignment.

done with the ordinary quick-to-release relays, for slow release relays the armature is allowed to contact the core, finish to finish. When plane flat pole face surfaces are provided, it is expensive and difficult to insure in commercial practice that precise and uniform alignment of the pole face surfaces will obtain. Variations in the alignment of these two surfaces will cause variations in the closed magnetic circuit reluctance and consequently on the release time of the relay.

In Fig. 12 is shown a design where the necessity for holding the alignment of core and armature so precisely is not so great.⁹ A spherical surface of rather large radius is embossed on the front end of the armature, so that with commercial variations in alignment, the armature always presents a point on the surface of a sphere for contacting the flat surface of the core. Similarly, the legs of the armature where they pivot on the front ends of the hinge bracket are likewise embossed. The results of the effects of these structural differences on the closed circuit reluctance are shown in Fig. 13 for a design with flat surfaces and one with embossed surfaces. While it is true that with perfect alignment the relay with flat surfaces will give longer release times, it is apparent that as variations in alignment occur from time to time and from relay to relay, it will have larger variations in performance than the relay with the embossed surfaces. This is a feature which has proven of great value in the manufacture of slow release relays of reasonable time precision.

REFERENCES

1. C. Schneider, "Cellulose Acetate Filled Coils," *Bell Labs. Record*, **29**, p. 514, Nov., 1951.
2. W. C. Slauson, "Improved U, UA and Y Type Relays," *Bell Labs. Record*, **29**, p. 466, Oct., 1951.
3. B. F. Runyon, "Contacts for Crossbar Apparatus," *Bell Labs. Record*, **18**, p. 278, May, 1940.
4. P. W. Swenson, "Contacts," *Bell Labs. Record*, **27**, p. 50, Feb., 1949.
5. H. M. Knapp, "The UB Relay," *Bell Labs. Record*, **27**, p. 355, Oct., 1949.
6. L. H. Germer and F. E. Haworth, *J. Appl. Phys.*, **20**, p. 1085, 1949.
7. L. H. Germer, *J. Appl. Phys.*, **22**, p. 955, 1951.
8. R. M. Bozorth, *Ferromagnetism*, D. Van Nostrand Co., Inc., 1951.
9. F. A. Zupa, "The Y-Type Relay," *Bell Labs. Record*, **16**, p. 310, May, 1938.

Impedance Bridges for the Megacycle Range

By H. T. WILHELM

(Manuscript received August 19, 1952)

This paper reviews ac bridges developed for use in the Bell System for the measurement of impedance parameters, particularly at frequencies in the megacycle range. Three recent bridges designed for measuring networks and components for coaxial systems are described.

INTRODUCTION

The need during recent years for increased accuracy of impedance measurement in the megacycle range has led to advances in the art of bridge measurement. A particular stimulus has been the development of a new coaxial system, designated L-3, for transmitting over distances up to several thousand miles a continuous frequency band extending roughly from 0.3 to 8 megacycles per second. Such a system will be capable of providing on a single coaxial unit the combination of a single television channel and as many as 600 one-way telephone channels. The large loss inherent in transmitting this wide frequency band over the cable makes it necessary to provide an amplifier about every four miles, and these amplifiers and associated networks have created difficult measurement problems.

MEASUREMENT PROBLEMS

The measurement problems arise partly from the wide frequency band, approximately thirty times the minimum frequency. This makes equalization of the system for satisfactory transmission very difficult, particularly in transmitting a television signal which covers a frequency band equivalent to about a thousand telephone channels and which must be equalized for phase as well as loss.

Even more important, however, are the problems arising from the close spacing of the amplifiers, with the result that a transcontinental circuit requires up to a thousand amplifiers in its path. Departures in

individual transmission characteristics will produce cumulative errors, making it necessary to maintain close control over the manufacture and adjustment of all of these amplifiers and associated networks. This calls for networks of highly refined design and requires ancillary measurement facilities of greater precision than heretofore available at these higher frequencies.

The design of transmission networks to meet exacting requirements is a subtle art, embracing on the one hand the use of complex mathematical manipulation to produce theoretical networks having the desired loss and phase characteristics, and requiring, on the other hand, a down-to-earth knowledge of the properties of the actual components used including parasitic effects and interaction of the various elements when assembled into a network. To furnish this knowledge, to measure the component resistors, capacitors, inductors and transformers which are the building blocks of the networks, to evaluate the ever-present parasitic effects, to determine simplified circuit equivalents of the more complex components such as transformers, and to answer other questions too numerous to mention, measurements of impedance parameters — precise measurements — are required.

EXISTING BRIDGE TECHNIQUE

For measuring impedance and admittance parameters, that is R , L , C and G , suitable ac bridges, ordinarily simply designated as impedance bridges, have long held a high place in the Bell System because of their inherent reliability and precision, and their ability to cover a wide range of values. The development of many of the original bridges^{1, 2, 3, 4} for frequencies above the audio range stemmed from the needs of the earlier carrier systems. With this development came also analysis of shielding technique,⁵ standardization of capacitance,^{6, 7} and a systematic classification of bridge methods⁸ by J. G. Ferguson in 1933, in which bridges were grouped into two major types designated as ratio-arm and product-arm, respectively. Following this classification, combined impedance and admittance bridges were developed,^{9, 10} utilizing a single set of bridge standards for both kinds of parameters by changing the configuration of the bridge network. There have also been special purpose bridges^{11, 12, 13, 14} for use at audio and the lower carrier frequencies. More recently, coaxial impedance standards¹⁹ having values calculable from physical dimensions have been developed.

Bridges for frequencies above one-half megacycle were used in the Bell System as early as 1919,¹⁵ but relatively few bridges were built until the mid 1930's when new carrier systems required bridges in the

megacycle range. A ratio-arm bridge¹⁶ using external standards was developed for precise measurements up to three megacycles. Interconnection of bridge and standards using coaxial cords provided flexibility of configuration resulting in an admittance bridge for high impedances and a series-reactance bridge for medium impedances. These two bridge circuits are shown schematically in (a) and (b) of Fig. 1. A separate, self-contained Maxwell product-arm inductance bridge, shown schematically in Fig. 1c and illustrated in Fig. 2, was designed primarily for measuring low-impedance parameters up to one megacycle/sec. Inductance was measured using calibrated air capacitors, and resistance was measured by means of conductance decades employing wire-wound resistors. The bridge included a double-shielded coupling transformer and complete shielding not shown in the simplified schematic.

To show clearly the scope and inter-relation of these three bridge methods, it is helpful to plot their ranges on a Slonczewski reactance/frequency chart¹⁷ shown in Fig. 3. In this chart, the top frequency shown for the ratio-arm bridge is three megacycles, and for the Maxwell bridge is one megacycle, as these are considered boundaries

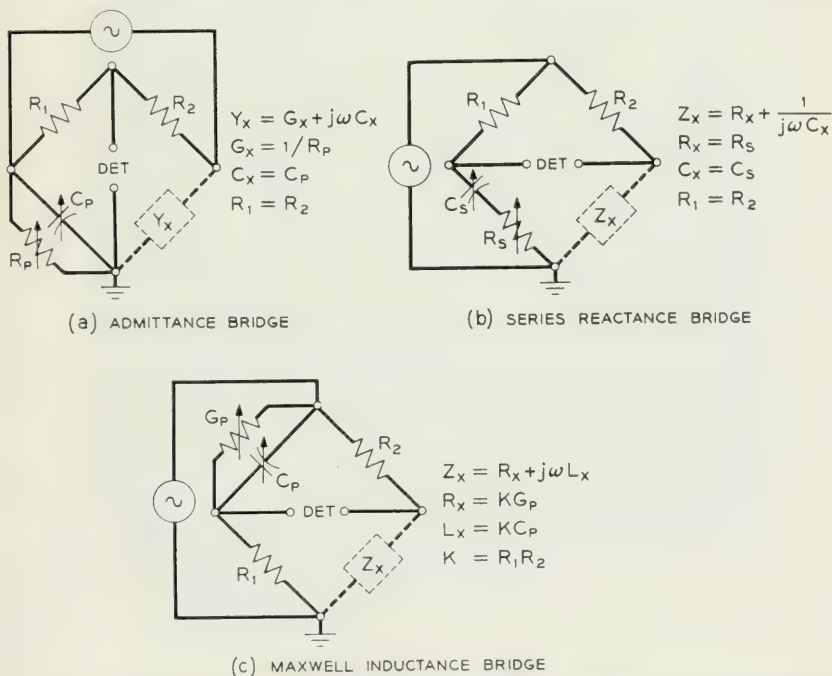


Fig. 1—Simplified schematics showing the basic circuits of three existing bridges for use at frequencies up to about three megacycles.

for their best performance, even though both bridges are useable at higher frequencies. It will be observed that while there is some overlapping of the three ranges, all three methods are necessary to obtain the impedance coverage shown. It should be emphasized that all the ranges shown cover both capacitive and inductive reactances. In the case of the admittance and series-reactance bridges, inductive impedances are measured by using a resonating capacitor, in parallel or series, respectively, with the apparatus being measured. In the Maxwell inductance bridge, capacitive impedances are measured by using a fixed resonating inductor in series with the impedance under test. A complete accuracy statement for these bridges is necessarily complex, but in general accuracies of ± 0.25 per cent for the major component

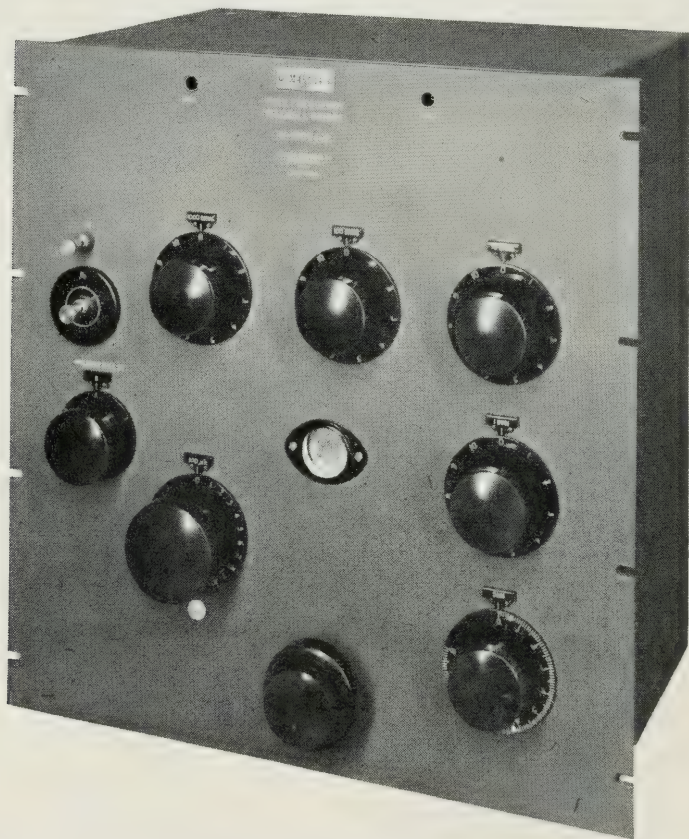


Fig. 2—One-megacycle Maxwell inductance bridge, shown schematically in Fig. 1c, designed for relay-rack mounting.

was obtained over most of the range plotted on the reactance chart. These bridges have been very successful for the purpose for which they were designed, but they are not useable up to the eight megacycles or higher required by the L3 system.

REQUIREMENTS OF BRIDGES FOR L3 SYSTEM

When the L3 system was contemplated, it was evident that new bridges would be needed. It was required to be able to measure virtually any impedance value at frequencies up to and beyond the second harmonic of the 8.4 megacycle upper limit of the system. Accordingly, a top

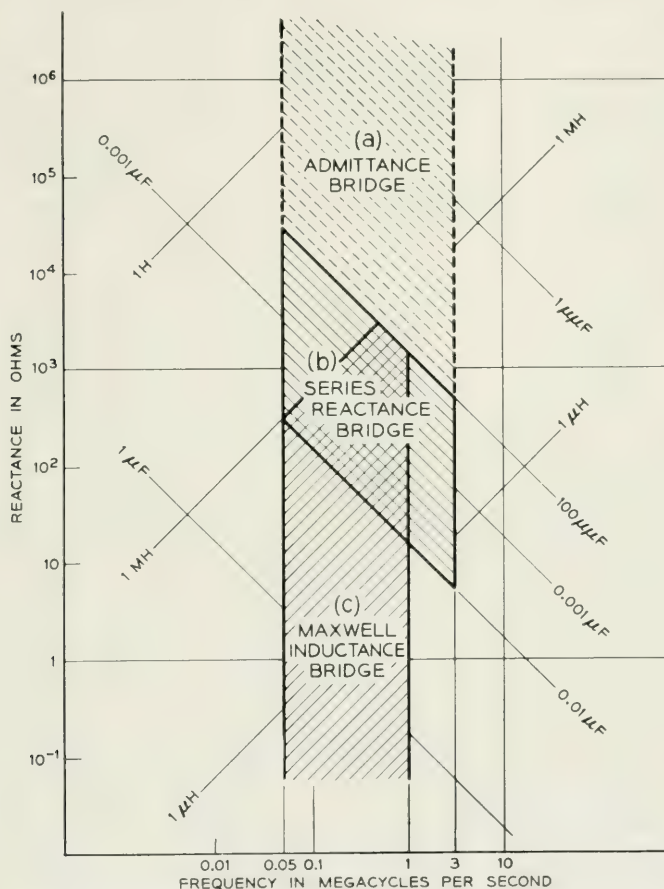


Fig. 3—Reactance/frequency chart showing the measurement range of the bridges shown in Fig. 1.

frequency of twenty megacycles was decided upon as a design objective with a basic accuracy of ± 0.5 per cent for the major component. The immediate need was for a general-purpose bridge, but it was expected that special-purpose bridges having better accuracy would be required later.

GENERAL PURPOSE 20-MEGACYCLE BRIDGE

It was decided first to develop a single bridge unit which would embrace both admittance and series impedance methods, and thereby cover a reactance range from a few ohms up to nearly a megohm, as shown in Fig. 4. Such a bridge would combine the features of (a) and (b) of Fig. 1. There were numerous departures from the earlier designs, however, including the use of a series range capacitor to reduce the size of the series capacitance standard, the use of deposited carbon resistors,¹⁸ the form and construction of both conductance and resistance standards, and especially the use of transformer-coupled inductive ratio arms.

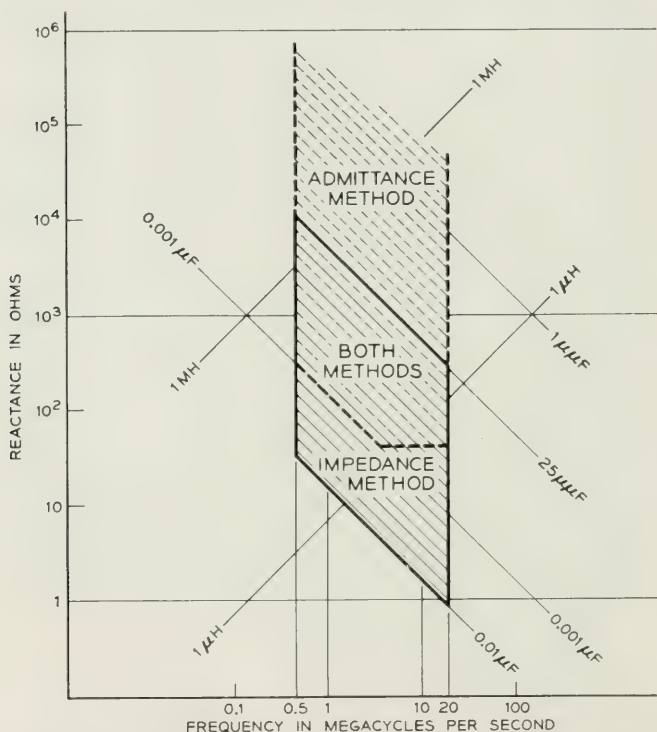


Fig. 4—Reactance/frequency chart applying to the general-purpose bridge shown in Fig. 5.

The successful use of a center-tapped transformer for ratio arms in a 465-KC direct capacitance bridge¹³ indicated that the resistance ratio arms r_1 , r_2 of Fig. 1 might be omitted if a suitable transformer could be developed for higher frequencies. The transformer group of the Laboratories succeeded in producing a transformer with a deviation from unity ratio of less than 0.1 per cent over a frequency range from 0.5 to 20 megacycles. This was made possible by precise location of the windings in fine milled grooves in the form of reversed helices, cut on a longitudinally-split brass cylinder for the inner winding, and on a surrounding phenol

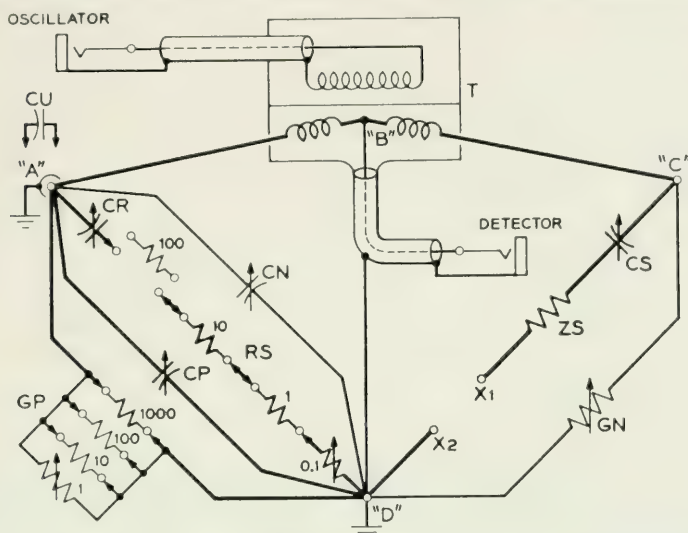


Fig. 5—Schematic of the 20-megacycle general-purpose bridge showing both the series (impedance) and parallel (admittance) bridge circuits combined in a single unit.

fibre cylinder for the bifilar outer winding which serves as the bridge ratio arms. Electrostatic shielding limits the direct capacitance between primary and secondary to less than $0.01 \mu\mu f$. The core material is compressed powdered molybdenum permalloy. This transformer was the nucleus around which the general purpose bridge was built, and the resulting bridge is shown schematically in Fig. 5.

In Fig. 5, the letters A, B, C and D designate the four bridge corners, and T is the ratio-arm transformer already described. Apparatus to be measured by the admittance method is connected to terminals C and D, and is balanced by the calibrated capacitor CP and conductance standard GP. To use the series reactance method, CP and GP are set at minimum settings, apparatus to be measured is connected to terminals X1 and X2,

and is balanced by the calibrated capacitor CS and resistance standard RS . The range capacitor CR consists of several mica capacitors for extending the range of CS , as will be described below; and ZS is merely a compensating impedance, essentially an inductive two-ohm resistor. The circuit is thus basically quite simple and avoids the use of switches or other complications which would impair performance at these high frequencies.

Capacitors CP and CS are worm-driven air capacitors with a range of about $220\ \mu\mu f$, and were specially designed for this bridge. In the case of CS , any direct conductance between rotor and stator would result in an effective series resistance which would vary both with frequency and capacitor setting, and therefore require laborious correction. This was avoided by arranging the construction so that the rotor and stator are mounted on independent insulating supports to the ground panel, thereby completely eliminating direct conductance from rotor to stator. While this results in some conductance from test terminal $x1$ to ground, the amount is small and its effect is negligible because of the relatively low impedance values measured. In the case of CP , on the other hand, it is important to minimize series resistance and inductance to avoid conductance and capacitance corrections which would change both with frequency and capacitor setting. This was accomplished by careful design of the rotor brush using silver contact surfaces and center-fed connections to both rotor and stator.

The conductance standard, GP , and resistance standard, RS , were designed to emphasize high-frequency performance. Deposited carbon resistors¹⁸ on ceramic rods $\frac{1}{8}$ " in diameter and $\frac{3}{4}$ " long mounted on small decade rotors were used, so arranged that only one resistor on a rotor is in the circuit at any time, and that adjacent resistors are short-circuited by means of auxiliary shorting brushes to eliminate shunting admittance which might vary with frequency. For GP the resistance values are such that the two lower decades and the slide-wire rheostat each have a residual conductance of 333 micromhos, thereby avoiding the use of resistors exceeding 3,000 ohms in value which would be more likely to vary with frequency. The structure is designed to minimize series inductance and to maintain constant capacitance for all settings. For RS , on the other hand, it is necessary to maintain constant inductance for all settings. This was accomplished by adding small wire-loop compensating inductors in series with individual resistors in the 10-ohm and 100-ohm decades when necessary. To minimize the over-all inductance, the resistor rotors are placed very close together and are driven by gearing from the corresponding dials.

The range capacitor, CR , has already been mentioned. It consists of a

rotor switch on which are mounted five uncalibrated mica capacitors which enable cs to measure both positive and negative reactance values up to 10,000 $\mu\mu f$ without additional switching. The 20 $\mu\mu f$ cr capacitor covers capacitance measurements up to 60 $\mu\mu f$; the 40 $\mu\mu f$ capacitor covers up to 150 $\mu\mu f$; the 80 $\mu\mu f$ up to 600 $\mu\mu f$; the 140 $\mu\mu f$ up to 10,000 $\mu\mu f$; and the 200 $\mu\mu f$ capacitor covers all the positive series reactance measurements. Since the cr capacitor permits the bridge to be balanced with the test leads short-circuited, the value of the effective resistance under test is simply equal to the difference between rs readings for the measurement balance and the short-circuit balance, and the reactance under test is determined from a computation of the two readings of cs .

A front view of the general purpose bridge is shown in Fig. 6. The four lower dials are for GP ; above them are the four rs dials; and above them is the cr dial. The capacitors cs and cp are located adjacent to the test terminals, but are operated remotely by the dial knobs at the extreme right end of the bridge. This was done to remove the operator's hands as far as possible from the test terminals. Near the test terminals is a coaxial connector engraved λ . This allows plug-in capacitors (cv in Fig. 5) to be added in parallel with cp for extending the capacitance range. Compact silvered mica capacitors in steps of 200 $\mu\mu f$ are used. Fig. 7 shows the interior of the same bridge with cp and cs in the lower foreground, GP at the left and rs in the upper right.



Fig. 6—Front view of the general-purpose bridge shown in Fig. 5. The bridge is approximately 10½ inches high and 19 inches wide.

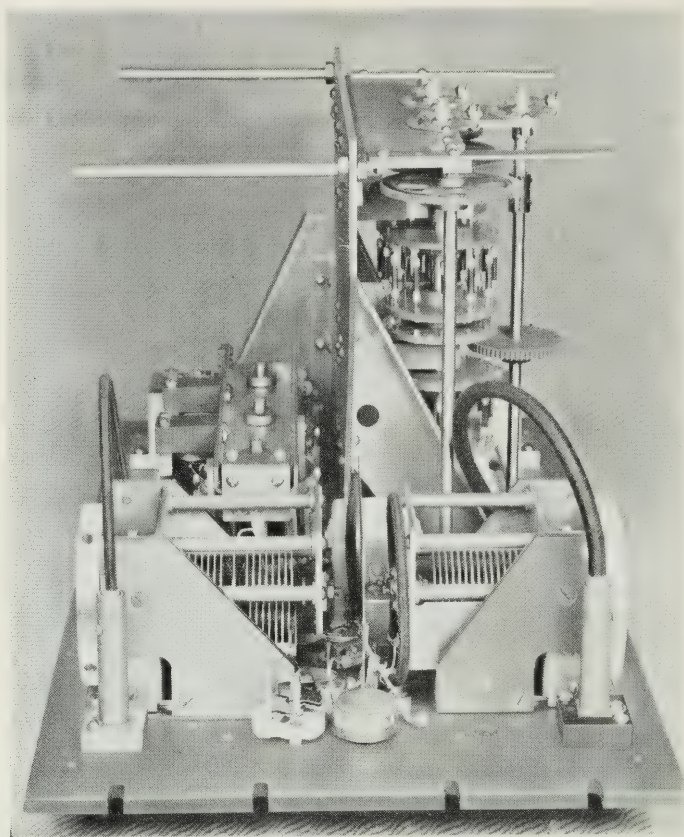


Fig. 7—Interior view of the general-purpose bridge. The panel edge shown in the foreground is the left edge of the bridge shown in Fig. 6.

FIVE-MEGACYCLE MAXWELL INDUCTANCE BRIDGE

To facilitate the measurement of low-valued inductors, there was need for a direct-reading inductance bridge inasmuch as such measurements entail considerable computation effort when using the general-purpose bridge. Accordingly, it was decided to build a five-megacycle Maxwell inductance bridge to cover a range from 0.001 microhenry up to 10 microhenries, and effective resistance values up to 11 ohms. The basic circuit is the same as the Maxwell bridge in Fig. 1, but the design embraces such refinements as glass-sealed deposited-carbon resistors for the conductance standard, and a worm-driven center-fed variable air capacitor. Special woven-wire resistors on spools of Teflon are used for the two fixed arms, and are compensated to give a constant product of practically zero

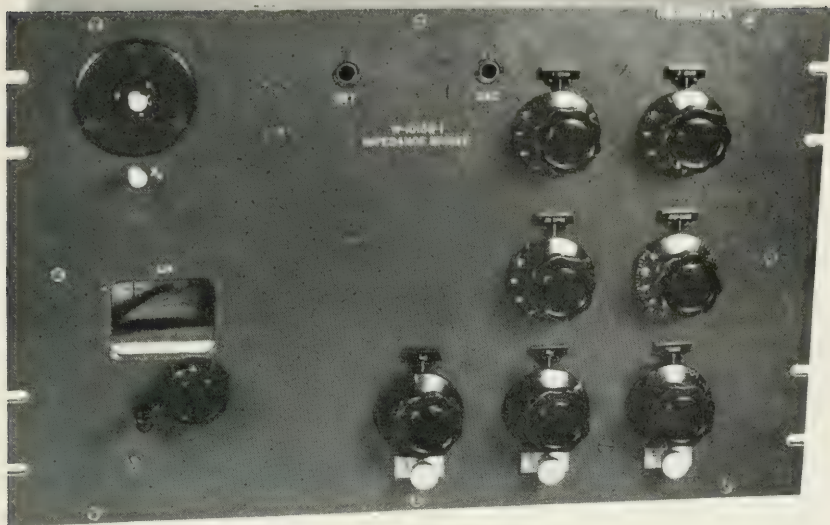


Fig. 8—The five-megacycle Maxwell inductance bridge is approximately $12\frac{1}{4}$ inches high and 19 inches wide. Test terminals are at upper left, and the three knobs at lower right are zero-balance adjusters.

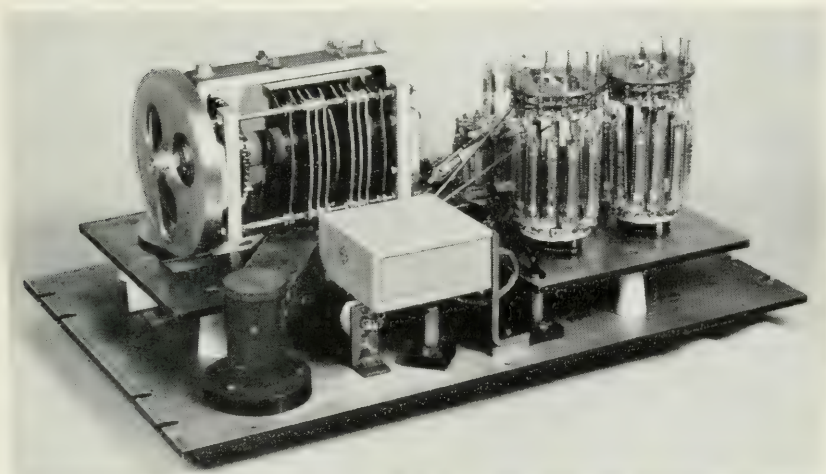


Fig. 9—Interior of bridge of Fig. 8 showing the shielding for test terminal X1 in foreground; at the left is the calibrated air capacitor; at the right are the conductance decades using glass-sealed deposited-carbon resistors.

phase angle over the entire frequency range. The result is a direct-reading bridge shown in Figs. 8 and 9 which has greatly facilitated the development of inductors in the megacycle range. The accuracy for major component varies from ± 0.25 per cent at one megacycle up to ± 1 per cent at five megacycles.

TEN-MEGACYCLE ADMITTANCE BRIDGE

Development of capacitors for the L3 coaxial system has required a new ten-megacycle admittance bridge. Intended especially for determining temperature coefficient and frequency characteristics of small capacitors, the bridge is capable of measuring capacitance values up to 200 $\mu\mu f$ with a precision of $\pm 0.01 \mu\mu f$, and a wide range of conductance values. Unlike the other two bridges described which make grounded measurements only, this bridge is arranged for direct and balanced-to-ground measurements as well. This is accomplished by using the ratio-arm transformer already described in combination with a simple grounding circuit using a three-position key, as shown in the bridge schematic of Fig. 10.

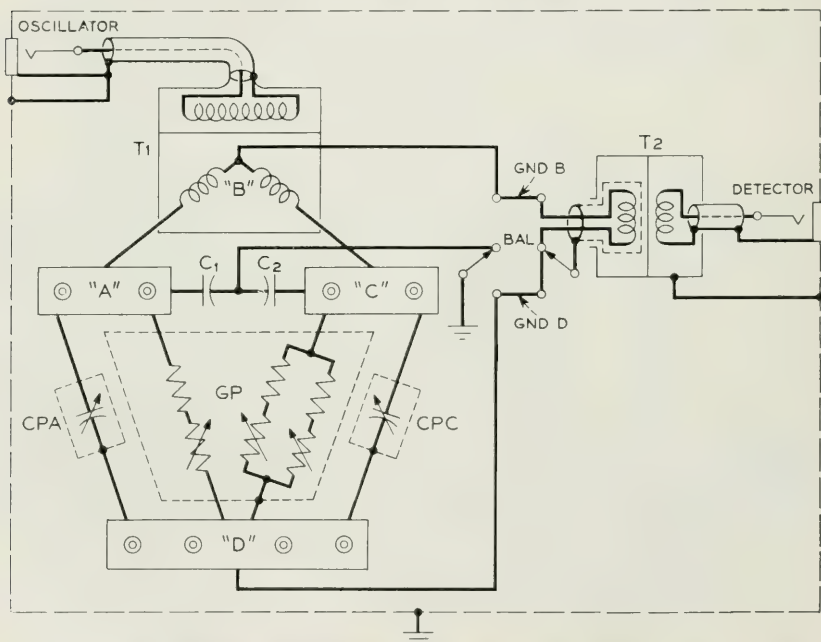


Fig. 10—Ten-megacycle admittance bridge with three-position key for shifting ground to B for measuring direct admittance, to junction of C1 and C2 for balanced admittance, and to D for grounded admittance. Unknowns may be connected from A to D or C to D.

The direct-capacitance measurements are useful in the development of low-valued capacitors, and the balanced-to-ground measurements are helpful in evaluating low-admittance off-ground networks.

CONCLUSION

Bridges have been developed for the measurement of impedance and admittance parameters at megacycle frequencies with accuracies heretofore possible only at much lower frequencies. Several of the twenty-megacycle general-purpose bridges have been built and are furnishing useful measurements of networks and components. Experience with these bridges has indicated ranges for which supplementary special-purpose bridges would be desirable, and two such bridges have been built: a Maxwell bridge for low-valued inductors, and an admittance bridge for low-valued capacitors. One feature of all of these bridges not generally available in commercial measuring instruments for megacycle frequencies is the provision of standards having a range of several decades. These allow balances to be made with greater precision over a wider range of phase angles in the apparatus under test, and assure that the absolute accuracy will not be limited by readability. This added precision is very useful in comparing similar components or in measuring characteristics such as temperature coefficient.

ACKNOWLEDGMENTS

The development of impedance bridges during the past thirty years has been under the direction of J. G. Ferguson, and the work described in this article has been under the supervision of S. J. Zammataro. Their assistance in the preparation of this paper has been very helpful and is hereby acknowledged. It is a pleasure also to acknowledge the contributions of a number of the author's colleagues particularly J. E. Nielsen who was largely responsible for the twenty-megacycle general-purpose bridge, and L. E. Herborn for the five megacycle Maxwell bridge.

REFERENCES

1. W. J. Shackelton, "A Shielded Bridge for Inductive Impedance Measurements," *Bell System Tech. J.*, **6**, pp. 142-171, Jan., 1927.
2. W. J. Shackelton and J. G. Ferguson, "Electrical Measurement of Communication Apparatus," *Bell System Tech. J.*, **7**, pp. 70-89, Jan., 1928.
3. J. G. Ferguson, "Measurement of Inductance by the Shielded Owen Bridge," *Bell System Tech. J.*, **6**, pp. 375-386, July, 1927.
4. S. J. Zammataro, "Impedance Bridges," *Bell Labs. Record*, **8**, pp. 167-170, Dec., 1929.
5. J. G. Ferguson, "Shielding in High-Frequency Measurement," *Bell System Tech. J.*, **8**, pp. 560-575, Aug., 1929.

6. J. G. Ferguson and B. W. Bartlett, "The Measurement of Capacitance in Terms of Resistance and Frequency," *Bell System Tech. J.*, **7**, pp. 420-437, July, 1928.
7. W. D. Voelker, "An Improved Capacitance Bridge for Precision Measurements," *Bell Labs. Record*, **20**, pp. 133-137, Jan., 1942.
8. J. G. Ferguson, "Classification of Bridge Methods for Measuring Impedances," *Bell System Tech. J.*, **12**, pp. 452-468, Oct., 1933.
9. S. J. Zammataro, "An Inductance and Capacitance Bridge," *Bell Labs. Record*, **16**, pp. 341-346, June, 1938.
10. H. T. Wilhelm, "Impedance Bridge with a Billion-to-One Range," *Bell Labs. Record*, **23**, pp. 89-92, Mar., 1945.
11. H. T. Wilhelm, "Measuring Inductance of Coils with Superimposed Direct Current," *Bell Labs. Record*, **14**, pp. 131-135, Dec. 1935.
12. H. T. Wilhelm, "A Bridge for Measuring Core Loss," *Bell Labs. Record*, **19**, pp. 92-96, Nov. 1940.
13. C. H. Young, "Measuring Inter-Electrode Capacitances," *Bell Labs. Record*, **24**, pp. 433-438, Dec., 1946.
14. H. T. Wilhelm, "Maxwell Bridge for Measuring Loading Coils," *Bell Labs. Record*, **28**, pp. 453-457, Oct., 1950.
15. Carl Englund, "Note on Radio Frequency Measurements," *Proc. Inst. Radio Engrs.*, **8**, pp. 326-333, Aug., 1920.
16. C. H. Young, "A 5-Megacycle Impedance Bridge," *Bell Labs. Record*, **15**, pp. 261-265, Apr., 1937.
17. T. Slonczewski, "A Versatile Nomagram for Circuit Problems," *Bell Labs. Record*, **10**, pp. 71-73, Nov., 1930.
18. R. O. Grisdale, A. C. Pfister, W. van Roosbroeck, Pyrolitic Film Resistors—Carbon and Borocarbon," *Bell System Tech. J.*, **30**, pp. 271-314, Apr., 1951.
19. R. A. Kempf, "Coaxial Impedance Standards," *Bell System Tech. J.*, **30**, pp. 689-705, July, 1951.

Abstracts of Bell System Technical Papers* Not Published in This Journal

A Full Automatic Private-Line Teletypewriter Switching System. W. M. BACON¹ and G. A. LOCKE¹. *Trans. A.I.E.E.*, **70**, Part 1, pp. 473-480, 1951. (Monograph 1837).

This paper describes a full automatic teletypewriter message switching system for use in private-line networks involving one or more switching centers and a multiplicity of local or long-distance lines, each of which may have one or more stations. This system provides fast teletypewriter communication from any station to any other station or group of stations in the network. At its point of origin a message first is perforated in tape accompanied by suitable directing and end-of-message characters, thereafter it is transmitted automatically, stored temporarily in perforated tape at a switching office, and then routed at high speed to its point or points of destination. Important features are the arrangements provided to permit efficient use of long full duplex transmission lines, the full automatic handling of multiple-address messages with only a single originating transmission, and the various guards and alarms which are provided to protect against loss of messages in case of trouble.

Operational Study of a Highway Mobile Telephone System. L. A. DORFF¹. *Trans. A.I.E.E.*, **70**, Part 1, pp. 31-37, 1951. (Monograph 1838).

The Dynamics of the Middle Ear and Its Relation to the Acuity of Hearing. H. FLETCHER¹. *J. Acoust. Soc. Am.*, **24**, pp. 129-131, March, 1952.

The transformer action of the middle ear as measured by Bekésy is shown to be the principal cause for the low acuity of hearing for low frequencies. Because of the very low mechanical impedance across the basilar membrane at low frequencies, large acoustical pressures in front of the ear drum produce appreciable acoustical pressures across the basilar membrane. For example, at 100 cps this pressure is thirty times and at 6000 cps it is one-tenth that created across the basilar membrane.

Diffusion of Donor and Acceptor Elements Into Germanium. C. S. FULLER¹. *Phys. Rev.*, **86**, pp. 136-137, April 1, 1952.

* Certain of these papers are available as Bell System Monographs and may be obtained on request to the Publication Department, Bell Telephone Laboratories, Inc., 463 West Street, New York 14, N. Y. For papers available in this form, the monograph number is given in parentheses following the date of publication, and this number should be given in all requests.

¹ Bell Telephone Laboratories

A Submarine Telephone Cable with Submerged Repeaters. J. J. GILBERT¹. *Trans. A.I.E.E.*, **70**, Part 1, pp. 564-572, 1951. (Monograph 1815).

Physical Structure and Magnetic Anisotropy of Alnico 5. Part I. R. D. HEIDENREICH¹ and E. A. NESBITT¹. *Jl. Appl. Phys.*, **23**, pp. 352-371, March, 1952. (Monograph 1976).

It is concluded from electron metallographic results that the high coercive force and anisotropy of Alnico 5 are caused by a very finely divided precipitate produced by the permanent magnet heat treatment. This precipitate is a transition structure rich in cobalt and is face-centered cubic with $a_0 = 10\text{\AA}$ and appears as rods growing along the [100] directions of the matrix crystal when no magnetic field is applied during heat treatment. The size of the precipitate rods at optimum properties is approximately 75-100 \AA by 400 \AA long. The spacing between rows of rods is about 200 \AA . The rods are not distinctly resolved in the electron images unless they are grown by aging at 800°C. Their orientation and structure is clearly evident in the electron diffraction patterns at all stages of growth. The precipitate responds to a magnetic field applied during heat-treatment both by suppression of nuclei making an angle greater than about 70° with the field and by the forcing of the rods off the [100] direction into that of the field. The precipitate rods tend to scatter in direction about the field vector when the field is off the [100] but are aligned accurately when the field is along [100].

Energy of a Bloch Wall on the Band Picture. I. Spiral Approach. C. HERRING¹. *Phys. Rev.*, **85**, pp. 1003-1011, March 15, 1952.

It is shown that the band or itinerant electron model of a solid is capable of accounting for the "exchange stiffness" which determines the properties of the transition region, known as the Bloch wall, which separates adjacent ferromagnetic domains with different directions of magnetization. In this treatment the constant spin function usually assigned to each running electron wave is replaced by a variable spin function. At each point of space the spin of a moving electron is inclined at a small velocity-dependent angle to the mean spin direction of the other electrons, and this gives rise to an exchange torque which makes the spin direction of the given electron precess as it moves through the transition region, the precession rate being just sufficient to keep it in approximate alignment with the macroscopic magnetization. Physical insight into the mechanisms involved is provided by a rigorous solution of the wall problem for a ferromagnetic free electron gas in the Slater-Fock approximation, although it is known that the free electron gas is not likely to be ferromagnetic in higher approximations. Rough upper limits to the exchange stiffness constants for actual ferromagnetic metals can be calculated without using any empirical constants other than the saturation moment and the lattice constant. The results are only a few times larger than the observed values.

Elastic and Plastic Properties of Very Small Metal Specimens. C. HERRING¹ and J. K. GALT¹. *Phys. Rev.*, **85**, pp. 1060-1061, March 15, 1952. (Monograph 1977).

¹ Bell Telephone Laboratories

A Scanner for Rapid Measurement of Envelope Delay Distortion. L. E. HUNT¹ and W. J. ALBERSHEIM¹. *Proc. I.R.E.*, **40**, pp. 454-459, April, 1952. (Monograph 1967).

A measuring device is described which instantaneously displays the envelope delay-frequency characteristic on a cathode-ray screen. Loop and one-way measurements of long-distance radio networks can be carried out. The frequency range extends from 60 to 80 megacycles; the limits of accuracy are 1 millimicro-second or 2 per cent of the measured delay range. Comparison of two characteristics can be carried out by superposition of alternate scanning traces.

The device has been found useful in measuring the delay distortion of the TD-2 radio-relay system and in designing and adjusting the delay equalizers needed to correct it.

Numerical Integration Near a Singularity. E. L. KAPLAN¹. *J. Math. Phys.*, **31**, pp. 1-28, April, 1952. (Monograph 1980).

Measurement of Diffusion in Semiconductors by a Capacitance Method. K. B. McAFEE¹, W. SHOCKLEY¹ and M. SPARKS¹. *Phys. Rev.*, **86**, pp. 137-138, April, 1952.

Probing the Space Charge Layer in a p-n Junction. G. L. PEARSON¹, W. T. READ¹ and W. SHOCKLEY¹. *Phys. Rev.*, **85**, pp. 1055-1057, March 15, 1952.

Control Methods Used in a Study of the Vowels. G. E. PETERSON¹ and H. L. BARNEY¹. *J. Acoust. Soc. Am.*, **24**, pp. 175-184, March, 1952. (Monograph 1982)

Relationships between a listener's identification of a spoken vowel and its properties as revealed from acoustic measurement of its sound wave have been a subject of study by many investigators. Both the utterance and the identification of a vowel depend upon the language and dialectal backgrounds and the vocal and auditory characteristics of the individuals concerned. The purpose of this paper is to discuss some of the control methods that have been used in the evaluation of these effects in a vowel study program at Bell Telephone Laboratories. The plan of the study, calibration of recording and measuring equipment, and methods for checking the performance of both speakers and listeners are described. The methods are illustrated from results of tests involving some 76 speakers and 70 listeners.

Current Multiplication in the Type-A Transistor. W. R. SITTNER¹. *Proc. I.R.E.*, **40**, pp. 448-454, April, 1952. (Monograph 1969).

One of the basic phenomena exhibited by transistors is current multiplication. In transistors of the point-contact type (one of these has been called the Type-A), the mechanism giving rise to this effect has been somewhat uncertain. Four

¹ Bell Telephone Laboratories

possible mechanisms of the current multiplication process in the Type-A transistor are discussed. One of the mechanisms is based on trapping holes in the collector barrier of the semiconductor. By means of this trapping model, the effect of emitter current and temperature on the current multiplication is predicted. It is shown that these predictions are in reasonable accord with experiment. Furthermore, assuming this model to hold, the trap density and activation energy (produced by forming) may be evaluated.

Faraday Rotation of Guided Waves. H. SUHL¹ and L. R. WALKER¹. *Phys. Rev.*, **86**, pp. 122-123, April 1, 1952.

Transistor Forming Effects in n-Type Germanium. L. B. VALDES¹. *Proc. I.R.E.*, **40**, pp. 445-448, April, 1952. (Monograph 1969).

Some of the effects of electrical forming of the collector of an n-type germanium transistor are discussed. Evidence is presented for the existence of a region of p-type germanium underneath the formed electrode, together with some indication of the size of the formed region. These experiments lend support to the p-n hook mechanism in that they explain the observed high values of alpha in transistors. This relation is discussed.

Domain Structure of Perminvar Having a Rectangular Hysteresis Loop. H. J. WILLIAMS¹ and M. GOERTZ¹. *Jl. Appl. Phys.*, **23**, pp. 316-323, March, 1952. (Monograph 1985).

An investigation has been made of the magnetic domain structure of Perminvar (43 per cent Ni, 34 per cent Fe, 23 per cent Co) ring specimens having rectangular hysteresis loops after heat-treatment in a magnetic field. Domain patterns obtained with colloidal magnetite showed curved domain boundaries extending completely around the rings, forming circles concentric with them. Changes in magnetization occur when an applied field causes the circular boundaries either to expand or contract so that there is a change in the relative values of clockwise and counter-clockwise flux. A nucleus of reversed magnetization was formed by making a small notch in a specimen, and this decreased the coercive force and hysteresis loss by a factor of two. It was found that in a 180° domain boundary it was possible to make the change in spin orientations, which occurs in going from one side of the boundary to the other, have either a right- or left-hand screw relation, by the application of a field of appropriate sign perpendicular to the surface. The effect of superposing an applied alternating field was also investigated, and an effective permeability of 4,000,000 was obtained.

Measuring Techniques for Broad-Band. Long-Distance Radio Relay Systems. W. J. ALBERSHEIM¹. *Proc. I.R.E.*, **40**, pp. 548-551, May, 1952. (Monograph 1971).

Line-up and maintenance of radio relay systems require sensitive yet rapid measurements. These are obtained by scanning the systems response as functions of time, frequency, and amplitude. Parameters thus scanned include the

¹ Bell Telephone Laboratories

transient response to step functions; frequency characteristics of gain, phase, impedance and their frequency derivatives; and amplitude characteristics of output nonlinear and of intermodulation products.

Aluminum Die Castings—The Effect of Process Variables on Their Properties. W. BABINGTON¹ and D. H. KLEPPINGER⁴. *Proc. A.S.T.M.*, **51**, pp. 169–197, 1951.

Diffusion in Alloys and the Kirkendall Effect. J. BARDEEN¹ and C. HERRING¹. pp. 261–288 of *Imperfections in Nearly Perfect Crystals*, Wiley N. Y., 1952, 490 p. Edited by W. Shockley, J. H. Hollomon, R. Maurer and F. Seitz. Symposium held at Pocono Manor, Oct. 12–14, 1950, by Committee on Solids, National Research Council.

Lightning Protection for Fixed Radio Stations. D. W. BODLE¹. *Tele-Tech*, **11**, pp. 58–60, 126+, June, 1952.

Common grounds, parallel conducting paths, and discharge gaps provide three important means for avoiding equipment damage from high current surges. Protection of connecting facilities must also be considered to preserve service.

Compression Tests on Lead Alloys at Extrusion Temperatures. G. M. BOUTON¹ and G. S. PHIPPS¹. *Proc. A.S.T.M.*, v. **51**, pp. 761–770, 1951.

Load-deflection measurements made during compression tests on lead and lead-alloy cylinders at various temperatures show the effects of alloying ingredients on the force required to produce deformation. The curves also furnish clues as to changes taking place in the materials during the course of the test. The load, P , to produce definite small deformation in pure lead at various temperatures, T , are shown to follow the relationship $P = Ae^{-BT}$, where A and B are constants for the material. This is the same relationship found by others in extrusion studies. The elements added to lead were those most commonly used in the manufacture of cable sheath, namely, antimony, arsenic, bismuth, silver, tellurium, and tin. The results show that the stronger alloys now used for cable sheathing deform less readily at extrusion temperatures than pure lead or the weaker alloys.

RF Phase Control in Pulsed Magnetrons. E. E. DAVID, JR.¹. *Proc. I.R.E.*, **40**, pp. 669–685, June, 1952.

This paper describes the behavior of a magnetron oscillator started in the presence of an externally applied rf exciting signal whose frequency is not greatly different from the unperturbed steady-state frequency of the magnetron.

Effect of Prior Strain at Low Temperatures on the Properties of Some Close-Packed Metals at Room Temperature. W. C. ELLIS¹ and E. S. GREINER¹. *J. Metals*, **4**, pp. 648–651, June, 1952. (Monograph 1966).

¹ Bell Telephone Laboratories

⁴ Frankford Arsenal, Philadelphia, Pa.

The Fatigue Test as Applied to Lead Cable Sheath. G. R. GOHN¹ and W. C. ELLIS¹. *Proc. A.S.T.M.*, **51**, pp. 721-740, 1951.

This paper discusses the more important factors affecting the design of laboratory test methods suitable for obtaining significant fatigue data from reversed bending tests on cantilever-beam specimens of lead cable sheathing alloys. Data are presented to show the effect of cycling rate, temperature, shape of specimen, alloy additions, and aging on fatigue life. The close correlation between bending fatigue tests on strip specimens and full size sections of cable is demonstrated. The fatigue data are analyzed in terms of (1) cycle life versus deflection, (2) cycle life versus strain, and (3) cycle life versus stress. Photomicrographs illustrating representative laboratory and field failures are included.

Thermal Conductivity of Germanium. A. GRIECO¹ and H. C. MONTGOMERY¹. *Phys. Rev.*, **86**, p. 570, May 15, 1952.

Bell System Cable Sheath Problems and Designs. F. W. HORN¹ and R. B. RAMSEY¹. *Trans. A.I.E.E.*, **70**, Part 2, pp. 1811-1816, 1951. (Monograph 1917).

Powdered Standards for Spectrochemical Analysis. E. K. JAYCOX¹. *Applied Spectroscopy*, **6**, pp. 17-19, May, 1952. (Monograph 1978).

Engineering for Low Product Cost and High Product Quality at the Western Electric Company. A. C. JONES³. *Ind. Quality Control*, **8**, pp. 53-59, May, 1952.

The Approximation with Rational Functions of Prescribed Magnitude and Phase Characteristics. J. G. LINVILL¹. References. *Proc. I.R.E.*, **40**, pp. 711-721, June, 1952.

A successive-approximations method is applied to the selection of network functions having desired magnitude and phase variation with frequency. The first approximation, the first set of pole and zero locations, can be selected on the basis of known solutions to similar problems or through use of a set of curves. In succeeding approximations the pole and zero locations are adjusted to decrease the deviation of the earlier approximations from the desired characteristics. The process adjusts the magnitude and phase characteristics simultaneously. Its flexibility permits accommodation of practical constraints not possible with other methods.

The Magnetic Structure of Alnico 5. E. A. NESBITT¹ and R. D. HEIDENREICH¹. *Elec. Eng.*, **71**, pp. 530-534, June, 1952. (Monograph 1981).

In the investigation of Alnico 5, two problems arose. What is the mechanism which enables the alloy to respond to heat treatment in a magnetic field? What causes the alloy to have a high coercive force of 600 oersteds? The first problem has been solved and progress has been made toward solving the second.

¹ Bell Telephone Laboratories

³ Western Electric Company

Single-Frequency Signaling System for Supervision and Dialing Over Long-Distance Telephone Trunks. N. A. NEWELL¹ and A. WEAVER¹. *Trans. A.I.E.E.*, **70**, Part 1, pp. 489–494, 1951. (Monograph 1841).

The single-frequency signaling system for long-distance telephone trunks frees dial calls from the range and other limitations imposed by dc signaling methods. It uses alternating currents in the voice range as the signaling medium and so can be used with any trunk of any length or type of line facility which meets voice-transmission requirements. The signaling requirements, design problems, main features of the circuit and equipment arrangements, and the operation of this system are outlined in this paper. The system described is the first practical arrangement of its type satisfactorily to meet all the conditions of telephone service in the Bell Telephone System.

Experimental Information on Slip Lines. W. T. READ, JR.¹ pp. 129–151 of *Imperfections in Nearly Perfect Crystals*, Wiley, N. Y., 1952, 490 p. Edited by W. Shockley, J. H. Hollomon, R. Maurer and F. Seitz. Symposium held at Pocono Manor, Oct. 12–14, 1950, by Committee on Solids, National Research Council.

On the Geometry of Dislocations. W. T. READ, JR.¹ and W. SHOCKLEY¹. pp. 77–94 of *Imperfections in Nearly Perfect Crystals*, Wiley, N. Y., 1952, 490 p. Edited by W. Shockley, J. H. Hollomon, R. Maurer and F. Seitz. Symposium held at Pocono Manor, Oct. 12–14, 1950, by Committee on Solids, National Research Council.

A Servo System for Heterodyne Oscillators. T. SŁONCZEWSKI¹. *Trans. A.I.E.E.*, **70**, Part 1, pp. 1070–1072, 1951. (Monograph 1883).

A constant rate of progression of frequency of a motor-driven heterodyne oscillator is obtained by comparing its output with a frequency standard. The result is fed into a servo loop which drives the motor at the proper speed. When used in connection with a level recorder a linear frequency scale is obtained which is more accurate than the static calibration of the oscillator.

Metallic Rectifiers in Telephone Power Plants. D. E. TRUCKSESS¹. *Trans. A.I.E.E.*, **70**, Part 2, pp 1464–1467, 1951. (Monograph 1987).

Metallic rectifiers are a comparatively new means of converting power from alternating current to direct current. Most of the component apparatus used in the Telephone Systems operates with direct current while the normal power source is alternating current. Therefore a static device without expendable parts which is obtainable in small and large current capacity lends itself as a means for power conversion in telephone power plants.

¹ Bell Telephone Laboratories

Contributors to this Issue

A. B. CLARK, B.E.E., University of Michigan, 1911. A. T. & T. Co., 1911-34; Bell Telephone Laboratories, 1934-. Toll Transmission Development Engineer, 1929; Toll Transmission Development Director, 1934; Director of Transmission Development, 1935; Director of Systems Development, 1940; Vice President, 1944. Bell System Chairman of Joint Subcommittee on Development and Research of the Edison Electric Institute and Bell System since 1938. Since June, 1951, Mr. Clark has been in charge of coordinating all Bell System programs at the Laboratories. During World War II he served both as a consultant to and a member of various divisions of the Office of Scientific Research and Development. In 1944 he was appointed Consultant to the Secretary of War, and in connection with this work made trips to the European and Mediterranean theaters of operation. Member of I.R.E., Tau Beta Pi, Sigma Xi, and A.A.A.S. and Fellow of A.I.E.E. and the Acoustical Society of America.

J. R. FRY, M.E., Cornell University, 1915. Western Electric Company, 1915-25. Bell Telephone Laboratories 1925-. Mr. Fry has been Assistant Switching Apparatus Engineer in the Switching Apparatus Development Department since 1946. Except for the years 1941-45, when he worked on military projects, most of Mr. Fry's Bell System service has been devoted to the design and development of electromagnetically operated switching apparatus such as relays, switches, registers, and selectors. Member of Eta Kappa Nu.

H. C. MONTGOMERY, A.B., University of Southern California, 1929; M.A., Columbia University, 1933. Bell Telephone Laboratories, 1929-. Prior to the war, Mr. Montgomery was engaged in studies of hearing acuity and the analysis of speech sounds. His recent work in the transistor physics group has been concerned with fluctuation phenomena in semiconductors.

SAMUEL P. MORGAN, JR., B.S., California Institute of Technology, 1943; M.S., California Institute of Technology, 1944; Ph.D., California Institute of Technology, 1947. Bell Telephone Laboratories, 1947-. A

research mathematician, Dr. Morgan specializes in electromagnetic theory. He has been particularly concerned with problems of wave guide and coaxial cable transmission. Member of the American Physical Society, Tau Beta Pi, and an associate member of Sigma Xi.

W. H. NUNN joined the Home Telephone and Telegraph Company of Los Angeles in 1915. He became Plant Staff Engineer in 1927; Traffic Engineer in 1928; General Traffic Engineer, Oregon, 1935; General Traffic Engineer, Northern California and Nevada, 1940; Traffic Operations Engineer, Pacific Telephone and Telegraph Company, 1942; and General Traffic Manager, Northern California and Nevada, 1947. In July of 1949 he transferred to the American Telephone and Telegraph Company as Traffic Facilities Engineer, and since March of this year has been Assistant Chief Engineer.

H. S. OSBORNE, B.S., Mass. Inst. of Technology, 1908; Eng. D., Mass. Inst. of Technology, 1910; A. T. & T., Co., 1910-. Since joining the American Telephone and Telegraph Company in 1910, Mr. Osborne has been with the company continuously: as engineer in the Transmission and Protection Department until 1914; assistant to Transmission and Protection Engineer, 1914-1920; Transmission Engineer, 1920-1939; Operating Results Engineer, 1939-1940; Plant Engineer, 1940-1942; Assistant Chief Engineer, 1942-1943; and Chief Engineer from 1943 until his retirement in August of this year. During the war Mr. Osborne was Special Consultant in the office of the Secretary of War and a member of the Telegraph Committee, War Communications Board. In addition, he is a member of the Industry Advisory Council, Federal Specifications Board; of the Industry Advisory Committee for Supply Cataloging, Munitions Board; and of the Domestic Communications Industry Advisory Committee to N.P.A. For many years he has been active in the work of the A.I.E.E.: Chairman, Standards Committee, 1923-1926; member, Committee on Communications, 1931-1934; member, Edison Medal Committee, 1936-1943 and 1947-1952; Chairman, Committee on Award of Institute Prizes, 1936-1939; Chairman, Technical Program Committee, 1936-1939; member, Publication Committee, 1936-1939; Chairman, Special Committee on Institute Activities, 1936-1937; member, Committee on Planning and Coordination, 1936-1942, 1945-1946, and 1947-1949; member, Alfred Noble Prize Committee, 1937-1942; Chairman, Finance Committee, 1939-1942; President, 1942-1943; Chairman, Executive Committee, 1942-1943; member, Board of Directors and Executive Committee, 1942-1945; member, John Fritz Medal Board of Award, 1942-1946; Chairman, Board of Trustees, A.I.E.E. Retire-

ment System, 1944-1945; member, Hoover Medal Board of Award, 1945-1951; and Chairman, Board of Trustees of Volta Memorial Fund, 1949-. He also has been active in the American Standards Association. He was long a member of the Board of Directors, and was Chairman of the Standards Council from 1942-1945 and Vice President 1948-1951. Since 1949 he has been President of the U. S. National Committee of the International Electrotechnical Commission. He is a member of the Joint Conference Committee on Standards of the Department of Commerce and ASA, and Chairman of the U. S. N. C. Executive Council Subcommittee. He is Fellow of the American Institute of Electrical Engineers, Acoustical Society of America, American Physical Society, American Association for the Advancement of Science, and of the Institute of Radio Engineers; and is a member of the American Society for Engineering Education and of Tau Beta Pi.

J. J. PILLIOD, E.E. 1908, D.E. (Hon.) 1939, Ohio Northern University; A. T. & T. Co., 1908-. From 1910 until 1943 Mr. Pilliod was associated with the Long Lines Department and the General Engineering Department of the American Telephone and Telegraph Company. From 1914 to 1918 he was Division Plant Engineer in Chicago; 1918-1920, Engineer of Transmission, New York City; 1920-1941, Engineer in charge of Long Lines Engineering Department; and 1941-1943, General Manager of the Long Lines Department. In 1943 he assumed his present position as Assistant Chief Engineer of the American Telephone and Telegraph Company. From October 1942 to April 1943 he was Chief of Signal Section, Production Division, Army Service Force. He is a Fellow of the A.I.E.E. and is a Trustee of Ohio Northern University and of Vassar College.

F. F. SHIPLEY, B.S. in E.E., Purdue University, 1925. A. T. & T. Co., 1925-34; Bell Telephone Laboratories, 1934-. Since 1948, Mr. Shipley has been switching engineer in charge of planning large automatic switching systems, both local and toll. This includes panel, crossbar, and large step-by-step systems. Member of the A.I.E.E., Tau Beta Pi, and Eta Kappa Nu.

H. T. WILHELM, B.S. in E.E., Cooper Union, 1927; E.E., Cooper Union, 1936. Western Electric Company, 1922-24; Bell Telephone Laboratories, 1925-. Since joining the Laboratories Mr. Wilhelm's work has been with the Transmission Apparatus Development Department, where he has designed electrical measurement apparatus and developed test methods. Member of A.I.E.E. and Tau Beta Pi.

THE BELL SYSTEM

Technical Journal

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

VOLUME XXXI

NOVEMBER 1952

NUMBER 6

KARLSRUHE
PUBLIC LIBRARY
DEC 1 1952

A New General Purpose Relay for Telephone Switching Systems

ARTHUR C. KELLER 1023

Comparison of Mobile Radio Transmission at 150, 450, 900 and
3700 Mc

W. REA YOUNG, JR. 1068

Common Control Telephone Switching Systems

OSCAR MYERS 1086

Mathematical Theory of Laminated Transmission Lines—Part II

SAMUEL P. MORGAN, JR. 1121

Transistors in Switching Circuits

A. EUGENE ANDERSON 1207

Abstracts of Bell System Papers Not Published in this Journal

1250

Contributors to this Issue

1256

THE BELL SYSTEM TECHNICAL JOURNAL

ADVISORY BOARD

S. BRACKEN, *President, Western Electric Company*

F. R. KAPPEL, *Vice President, American Telephone
and Telegraph Company*

M. J. KELLY, *President, Bell Telephone Laboratories*

EDITORIAL COMMITTEE

E. I. GREEN, *Chairman*

A. J. BUSCH

F. R. LACK

W. H. DOHERTY

J. W. MCRAE

G. D. EDWARDS

W. H. NUNN

J. B. FISK

H. I. ROMNES

R. K. HONAMAN

H. V. SCHMIDT

EDITORIAL STAFF

PHILIP C. JONES, *Editor*

M. E. STRIEBY, *Managing Editor*

R. L. SHEPHERD, *Production Editor*

THE BELL SYSTEM TECHNICAL JOURNAL is published six times a year by the American Telephone and Telegraph Company, 195 Broadway, New York 7, N. Y. Cleo F. Craig, President; Carroll O. Bickelhaupt, Secretary; Donald R. Belcher, Treasurer. Subscriptions are accepted at \$3.00 per year. Single copies are 75 cents each. The foreign postage is 65 cents per year or 11 cents per copy. Printed in U. S. A.

THE BELL SYSTEM TECHNICAL JOURNAL

VOLUME XXXI

NOVEMBER 1952

NUMBER 6

Copyright, 1952, American Telephone and Telegraph Company

A New General Purpose Relay for Telephone Switching Systems

By ARTHUR C. KELLER

(Manuscript received July 14, 1952)

This paper describes a new general purpose electromagnetic relay for use in telephone switching systems. It is a wire spring relay known as the AF type relay and, with variations which provide slow release or marginal characteristics, it is known as the AG and AJ relay, respectively. Fig. 1 shows a typical AF type relay, Fig. 2 shows all of the parts of the relay and Fig. 3 is a drawing showing the relay assembly.

1. BACKGROUND

The general purpose relay is one of the most important components of telephone switching systems.¹ These relays constitute the most repetitive building block in switching equipment. Since several million are produced annually, low manufacturing cost is extremely desirable. Also of prime importance are low operating and maintenance costs. General purpose relays are, therefore, under constant observation and study by the telephone operating companies as the users, by the Western Electric Company as the manufacturer, and by Bell Telephone Laboratories as the designer. The AF wire spring relay and its variations are the result of such studies.

A general purpose relay for telephone switching systems must meet a large number of diverse requirements. It must be capable of being assembled with any one of a variety of magnet coils having a wide range

¹ S. P. Shackleton and H. W. Purcell, "Relays in the Bell System", *Bell System Tech. J.*, Jan., 1924, p. 1.

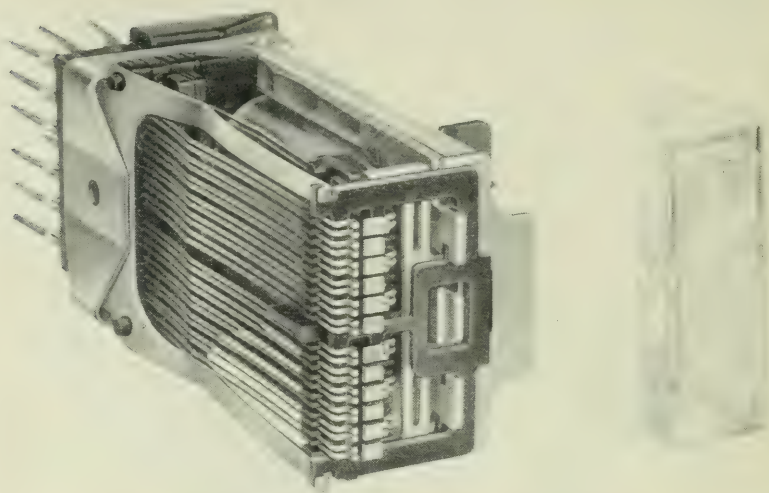


Fig. 1—AF type relay, with contact cover detached.

of resistance values and to operate contacts which vary from one pair to as many as fourteen or more. The basic relay design must also be capable of providing such features as fast operation and release, slow release, high sensitivity, heavy duty and marginal operation. These functions are performed satisfactorily in present crossbar switching systems by U, UA, UB and Y type relays.^{2, 3, 4, 5} However, with an objective of a forty-year life for new switching systems and a trend toward unattended operation of switching offices, it is important to attain the best in the performance and reliability of relays.

The general purpose relay must be designed to produce the best economic balance, when used in telephone switching systems, so that the annual charges are minimized. The major ingredients of these annual charges are manufacturing expense, operating electrical power, speed of operation and release, space required and maintenance costs which include reliability and life.

2. REQUIREMENTS AND OBJECTIVES

The requirements for a new general purpose relay were initially broadly stated to be performance and maintenance at least equal to the

² H. N. Wagar, "The U-Type Relay", *Bell Lab. Record*, May, 1938, p. 300.

³ H. M. Knapp, "The UB Relay", *Bell Lab. Record*, Oct., 1949, p. 355.

⁴ F. A. Zupa, "The Y-Type Relay", *Bell Lab. Record*, May, 1938, p. 310.

⁵ W. C. Slauson, "Improved U, UA and Y Type Relays", *Bell Lab. Record*, Oct., 1951, p. 466.

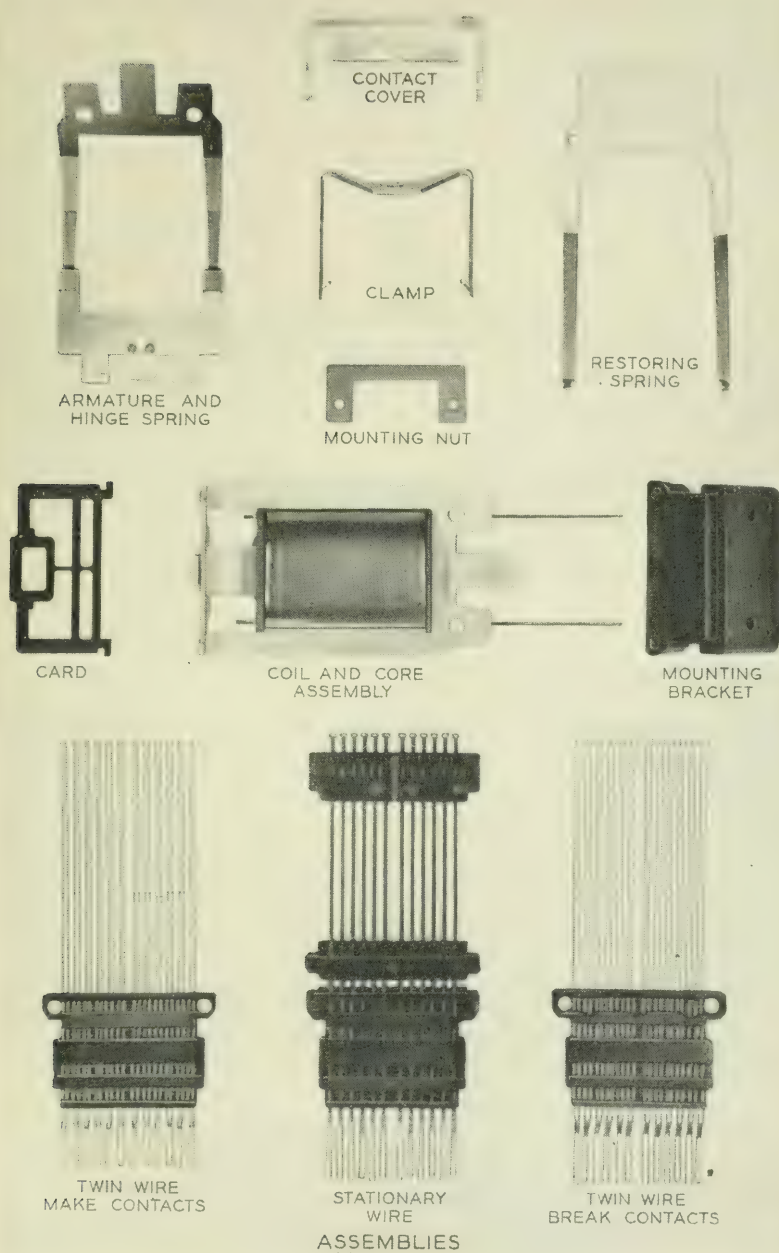


Fig. 2—Parts of the wire spring relay.

U and Y type relays but with substantially lower manufacturing costs. As the development of the relay proceeded, it became possible to expand the requirements without appreciably altering the expected relay cost. In particular, it became possible to design the new relay to operate and release faster or to use less electrical power, to operate more often before appreciable wear occurred, etc. The improved performance characteristics of the AF wire spring relay, as described later, are of equal economic importance to those associated with lower manufacturing cost.

The broad requirements were reduced to the following design objectives:

1. Lower cost—50 per cent of U type relay.
2. Reduced operating electrical power.
3. Faster operate and release times.
4. Long life—one billion operations.
5. Improved contact performance.

These broad design objectives do not specifically state a large number of other characteristics which must be at least as favorable as those of the U and Y relay family. This refers to such items as: space required, magnetic interference, wiring costs, contact combinations, field servicing and repairs.

3. DESIGN POSSIBILITIES

The studies of new relay design possibilities started with a careful review of the U type relay experience. In fact, much of the early thinking considered various modifications of U type and other existing relays. In general, these studies indicated that about half of the manufacturing cost of U type relays came from assembly and adjusting operations. Accordingly, these operations required major revision for a substantial

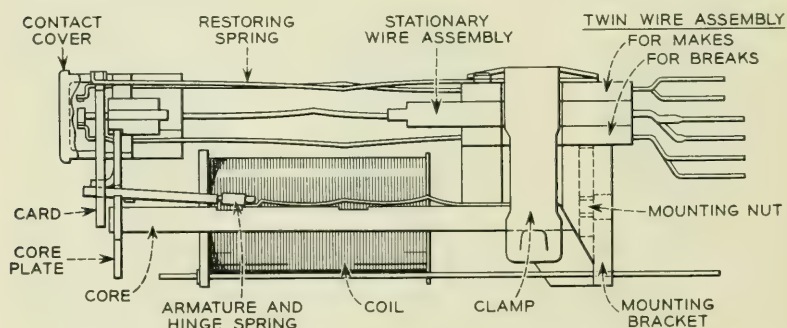


Fig. 3—Top view of the relay, showing location of parts.

cost reduction of the relay. It became evident that the development of new manufacturing methods as well as new designs were essential in reaching the ambitious objectives. For these reasons, the manufacturing engineers of the Western Electric Company were active participants in the development of the new relay from the beginning.

Many new forms of relay designs were considered and studied including such types as miniature, magnetic contact, piezoelectric, etc. As a result, one general form, first proposed by H. C. Harrison, gave the most promise of meeting the manifold requirements. This is the wire spring type characterized by the wire spring subassemblies with code card operation of pretensioned, low stiffness springs. Actually, the general form of the wire spring relay proposed by Mr. Harrison constitutes an entire new class of relays with many possible variations. These include various types of code card operation and various forms of contact operation, operated by any of a number of magnet structures.

The new class of relays has the following important advantages:

1. Pretensioned, low stiffness wire springs make possible (a) assembly to give close control of contact force without individual spring adjustment; and (b) essentially constant contact force throughout the life of the relay and its contacts.

2. Wire spring subassemblies make possible (a) favorable manufacture of a multiplicity of contact springs by molding; (b) lower assembly costs because fewer piece parts are needed; and (c) simple code card operation.

3. Code card operation makes possible (a) standardized and simple assembly; (b) accurate control of contact position; (c) essential elimination of locked contacts; (d) complete independence of twin contacts; and (e) simple means for providing a large number of contact combinations.

A continuous and comprehensive study was necessary of the characteristics and probable manufacturing costs of many forms of the wire spring relay family. As a result, after passing through several major designs, the basic design of the present relay was adopted. H. M. Knapp and C. F. Spahn proposed important features of this design. This form represented advantages over other types in

1. reducing the number and amount of dimensional variations controlling the contact gaps. In turn, this made possible smaller armature movement, shorter operating and release times and less chatter of the contacts;

2. reducing the number of code cards required to provide the large number of contact combinations needed in switching systems;

3. reducing the manufacturing and wiring costs;

4. increasing the mechanical life.

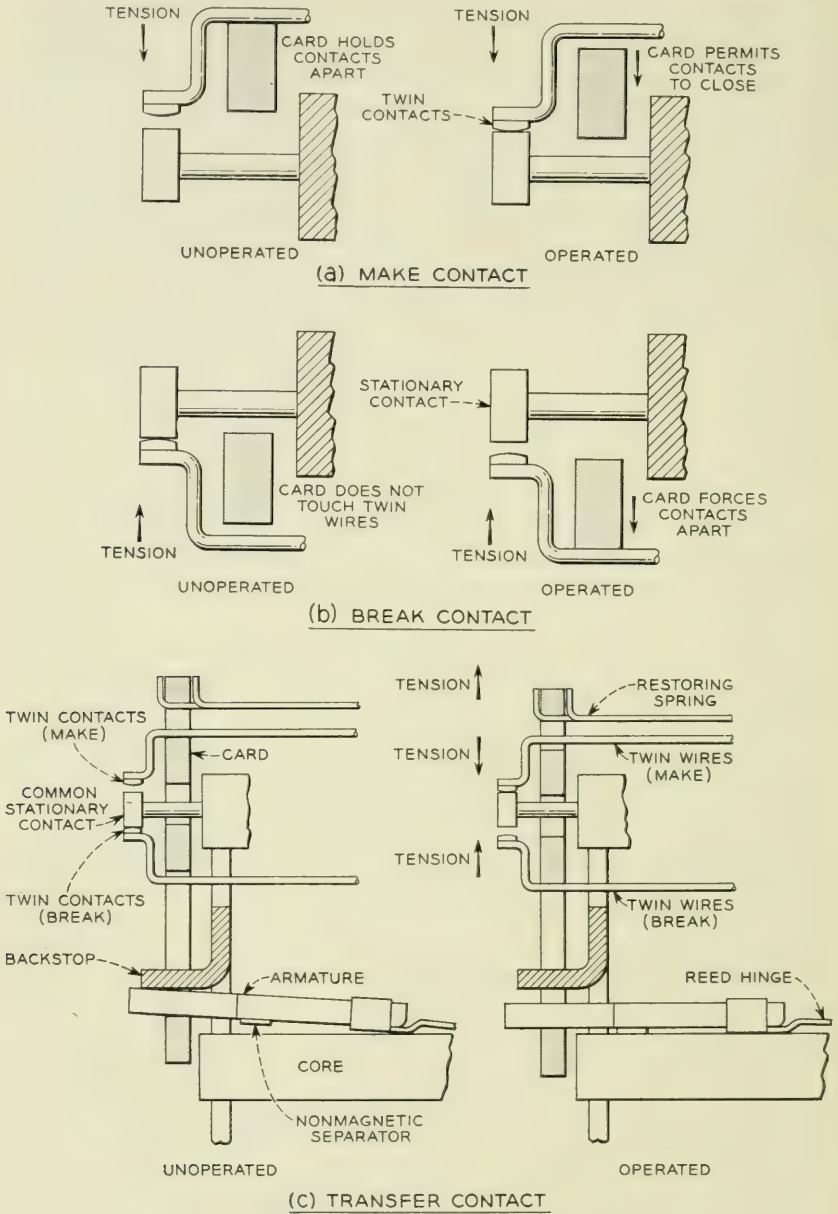


Fig. 4—Principle of contact operation.

4. PRINCIPLE OF CONTACT OPERATION OF THE AF RELAY

The AF relay uses what has been called the "single card system" for actuating the contacts. This is in contrast to other code card systems which require two, three or four coded cards in each relay. The method for obtaining individual make and break contacts with this system is shown in Figs. 4a and 4b, and a means for obtaining transfer contacts, in which both make and break twin contacts are associated with a common stationary contact, is shown in Fig. 4c. As indicated on the figures, the following principles are incorporated in this method of actuation:

1. In general, three basic wire spring assemblies are required. Two of these carry movable twin wires for make and break contacts and are identical except for some details in forming at the terminal ends for convenience in wiring. The twin wire assemblies are mounted on either

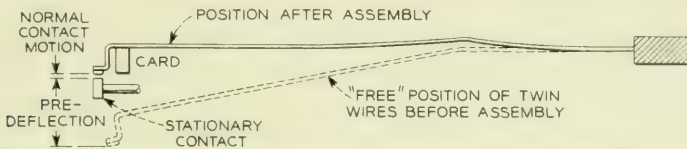


Fig. 5—Contact forces are controlled by relatively large predeflections of the twin wires.

side of the stationary wire assembly, which consists of a group of relatively heavy wires molded into plastic sections, one a short distance behind the contacts and one near the rear of the relay. These sections are rigidly supported in the relay structure.

2. Moving twin contacts on separate twin wires are used with every stationary contact. This arrangement assures good reliability and greater freedom from open contacts in the presence of dust and dirt. In addition, contact chatter is reduced as both contacts must be open simultaneously in order to interrupt the circuit.

3. As shown in Fig. 5, each group of twin wires is tensioned toward the stationary wires by means of large predeflections before assembly, so that the contact forces are determined by this predeflection. Good control of the contact force is assured without need for hand adjustment because small variations in deflection of the low stiffness springs do not result in appreciable changes in force. For this reason, the force is stable and is not appreciably affected by wear of the contacts.

4. The twin wires are actuated by a single punched fiber card. Since the tension in the twin wires is always in a direction to hold the contacts closed, the card serves to hold the make contacts open when the relay is unoperated and the break contacts open when the relay is energized.

5. The card is supported by the armature on one side and a restoring spring on the other. The restoring spring supplies the force to hold the armature against the backstop and to hold make contacts open when the relay is unoperated, while the armature supplies the force to hold the break contacts open when the relay is operated. However, since the armature must also overcome the tension in the restoring spring, the entire spring load must of course be overcome by the pull of the armature.

6. The twin contacts are held in good registration with their associated stationary contacts by means of molded guide slots in the stationary plastic member just behind the card. These guide slots are slightly wider than the diameter of the twin wires so that these wires are free to move in the direction of the armature movement, but are restrained against lateral motion.

7. The close proximity of the card to the contacts is important in minimizing contact chatter and in substantially eliminating locked contacts, i.e., contacts which fail to open because of interlocking of roughened surfaces. The close spacing results in a rigid coupling between the card and contacts, so that the static and dynamic forces associated with the armature and card are available to break loose any incipient lock which might develop.

As the armature moves toward the core, the particular point in its travel at which make contacts close and break contacts open depends upon the dimensions of the card between the surface which bears against the armature and the surfaces which engage the twin wires. By proper selection of these dimensions, any contact can be controlled to operate early or late in the travel as desired. By this means, several sequential contact arrangements may be obtained. For example, if the break contact in Fig. 4c is controlled by the card dimensions to open earlier in the travel than its associated make contact closes, the resulting arrangement is called an "early break-make" transfer. Similarly, an "early make-break" transfer, often called a "continuity" may be obtained by selection of card dimensions which will assure that the make contact closes before the break contact opens. If both contacts operate simultaneously, the result is a "non-sequence" transfer.

From the above it is evident that the card surfaces which engage the twin wires must be in different positions for early contacts as compared with late contacts. This is illustrated in Fig. 6 which shows an early break-make, an early make-break and a non-sequence transfer side by side. Of the contact pairs shown, only two operate early, and this is accomplished by means of steps in the actuating surfaces of the card.

Thus, if no sequences were required, the card would have a single straight surface for makes and another for breaks, and only one card variety would be needed for all combinations of makes, breaks, and non-sequence transfers. Where sequences are needed, however, additional card varieties are required with steps in the actuating surfaces for the early contacts.

In order to obtain a wider variety of the contact combinations including various numbers of make contacts, break contacts, sequence transfers and non-sequence transfers on the same relay, it is necessary to provide a variety of different coded stationary and twin wire assemblies, as well as a variety of cards, some of which are illustrated in Fig. 7. The twin wire assemblies differ as to the number of twin wires provided and in the position of these wires across the width of the molded section.

The stationary wire assemblies are always provided with a full complement of twelve wires in order to support the front molded section, which is held in place by spring tension in these wires. However, only certain of the wires may have contacts at the ends. These stationary contacts consist of base metal blocks with 0.010 inch thick precious metal surfaces on either or both sides as needed for makes, breaks or transfers, and any of the three varieties may be welded to any wire. Thus precious metal is provided only where needed for the particular contact arrangements desired.

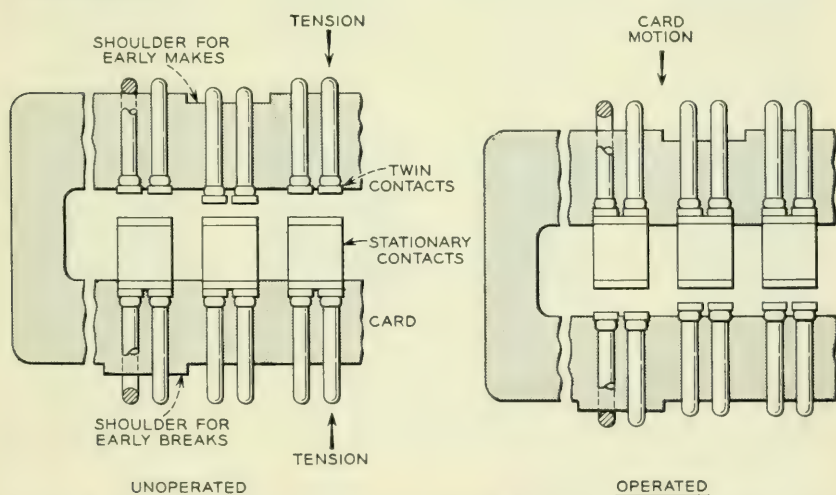


Fig. 6—Early break-make, early make-break and non-sequence transfer contacts, showing how early contacts are obtained by means of shoulders on the actuating card.

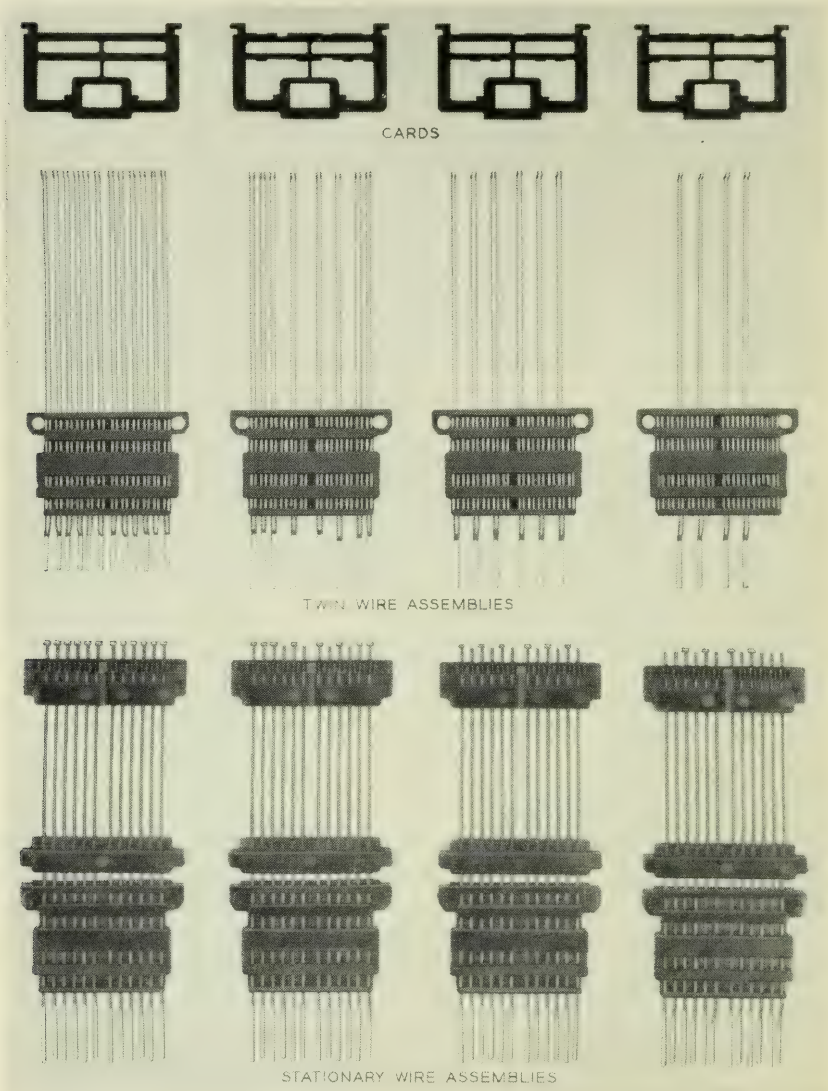


Fig. 7—A few varieties of the coded parts used to obtain various contact combinations.

By using different combinations of stationary and twin wire assemblies with each card variety, a large number of different contact combinations may be obtained. While most of these needed for telephone switching systems use either no sequences at all or a single stage of sequence, a few combinations are provided with "preliminary" contacts. These combinations include two stages of sequence, in which some contacts operate at each of three different points in the armature travel. The preliminary contacts operate earliest in the travel. These are followed by the early contacts of sequence transfers and finally by the late contacts, including ordinary makes and breaks.

To be sure the desired sequences will be maintained during the life of the relays, it is necessary to provide margins in the form of armature travel allowances at each stage. Combinations with sequences will therefore require total armature travels which are longer than those with no sequences, and two stages of sequence will require more travel than a single stage. Accordingly, the AF relay is provided with a choice of three armature travels to correspond with the number of sequences needed. At the card, these travels are 0.026 inch (short) for no sequences, 0.044 inch (intermediate) for one stage and 0.060 inch (long) for two stages.

Thus, combinations including ordinary makes, breaks and non-sequence transfers use short travel. Where sequence transfers are also needed, intermediate travel is used and the early contacts of the sequence transfers operate first. Long travel is used only where preliminary contacts followed by sequence transfers are needed.

5. ARMATURE SYSTEM AND MAGNETIC CIRCUIT

The armature system and the associated magnetic circuits constitute the basic motor element of an electromagnetic relay. The size of the motor element is determined, in part, by the work it must do and here a basic factor is the contact force. On the basis of analytical as well as experimental studies, it was decided to use a contact force of about six grams per single contact, i.e., about twelve grams for the combined force of the twin contacts. Other important factors which react on the design of the magnet are the speed required, winding space, heating,⁶ sensitivity, etc. The detailed analysis of the magnetic system and the associated measurements will be covered in separate papers.

The magnetic structure chosen is shown in Fig. 8. The armature is a flat member of U shape which provides desirably large pole face areas.

⁶ R. L. Peek, Jr., "Internal Temperatures of Relay Windings", *Bell System Tech. J.*, Jan., 1951, p. 141.

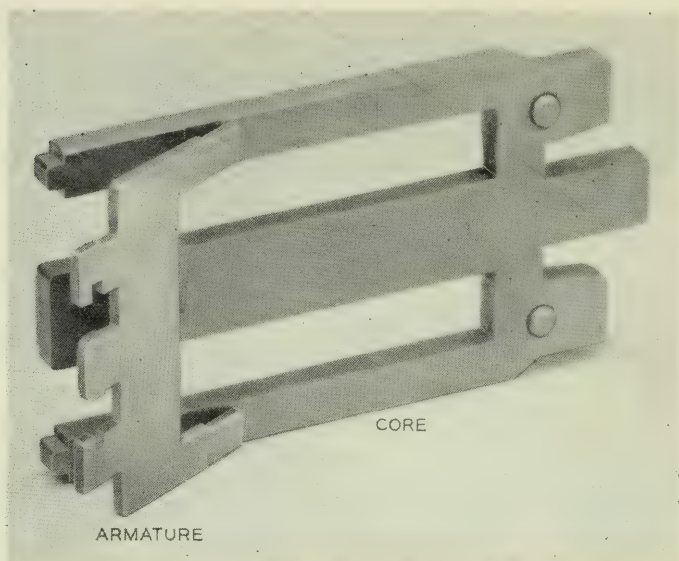


Fig. 8—Magnetic structure of the AF relay.

The core is a simple one-piece E-shape section of sufficient thickness of 1 per cent silicon iron to produce the magnetic flux needed to meet the force and speed requirements and to provide the main member to which all other parts are assembled. The silicon iron has appreciably higher electrical resistance than ordinary magnetic iron and this, together with the rectangular cross-sections of the legs, reduces eddy currents as needed for high speed operation and release. The one-piece construction avoids welded or butt joints common to many magnets. These joints are responsible for added reluctance and hence decrease the magnet sensitivity and require added electrical power to operate a given load. The relatively wide spacing of the legs increases leakage reluctance and, in turn, increases the useful magnetic flux.

After a cellulose acetate filled coil⁷ has been assembled to the middle leg of the core, a core plate, shown in Fig. 9, is forced over the ends of the E-shaped core to hold the three legs in good alignment for proper mating with the armature. The core plate also provides the backstop for the armature and serves as a means of gang adjustment of the contacts covered more completely under the Relay Adjustment section of this paper.

The armature is spring supported in a very definite manner to produce

⁷ C. Schneider, "Cellulose Acetate Filled Coils", *Bell Lab. Record*, Nov., 1951, p. 514.

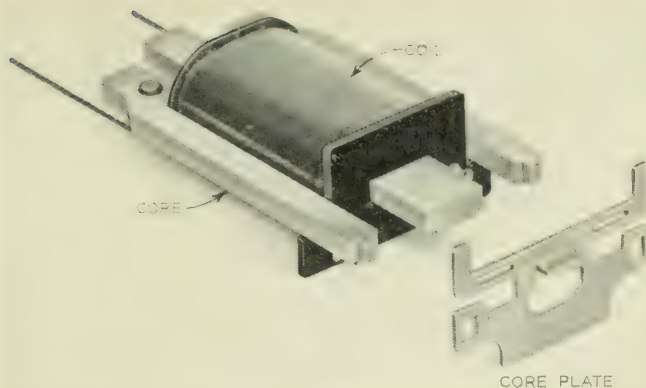


Fig. 9—Legs of the core are held in alignment by the core plate, which is forced over the ends after the coil is assembled.

a minimum of rebound when it is released from its operated position. The conditions for reducing armature rebound were described previously⁸ and make it necessary to proportion the forces at the front and rear of the armature properly. The magnitude and the ratio of these forces are a function of the mass distribution of the armature.

The magnet design must not only meet such functional requirements as speed, sensitivity, etc., but it must meet these for several values of armature travel as needed by the variety of contact combinations provided. Another requirement is that the relay be designed to fit on a 2-inch mounting plate and this, in turn, restricts the width of the E-shaped magnet core to slightly less than two inches. The relay is normally mounted with the 2-inch dimension in the vertical direction to allow the contact surfaces to be in vertical planes. The corresponding horizontal dimension in which the relay can be mounted is $1\frac{1}{2}$ inches except for a few special cases. As described in more detail under the section on Relay Performance, the improved magnet design has resulted in a reduction of the magnetic interference between mounted relays to values which are negligible for most practical purposes.

For comparison with the U type relay, the following typical constants of the magnet are of interest (see Table I). The closed gap reluctance, \mathcal{R}_0 , is the reluctance of the magnetic circuit, excluding leakage paths, with the armature operated and with the iron near maximum permeability. The coil constant, G , is the ratio of the square of the number of turns to the resistance for a full sized coil. The sensitivity, S , is a measure

⁸ E. E. Sumner, "Relay Armature Rebound Analysis", *Bell System Tech. J.*, Jan., 1952, p. 172.

of the ultimate work capacity of the magnet as related to the power input and has been defined as $S = 5\pi G/\mathcal{R}_0$ ergs per watt.

The favorable low value of closed gap reluctance for the new relay results from adequate cross-sections of magnetic material, the absence of joints, proper mating of the armature and core, and large pole face areas. A low value of reluctance also insures less magnetic interference to other relays and from other relays.

TABLE I

	AF Relay	U Relay
Closed Gap Reluctance \mathcal{R}_0 , cm^{-1}	0.028	0.065
Coil Constant G, kilomhos.....	160	160
Sensitivity S, ergs per watt.....	90,000	39,000

Although the coil constants are the same for the two relays, as can be seen from Table I, the sensitivity of the new relay is more than double that of the U relay, because of the lower closed gap reluctance.

6. MOLDED WIRE SPRING SUBASSEMBLIES

One of the major features of the new relay is the use of molded wire spring subassemblies. Fig. 10 shows a wire spring relay with twelve make contacts, and Fig. 11 shows a comparison of the wire spring assemblies used in this relay and the corresponding parts of the U relay. From this it is clear that the number of parts handled in the assembly of the contact spring members is greatly reduced in the new relay. Not all relays will have twelve contacts and in those cases where fewer contact springs are needed the comparison will not be so unfavorable to the U relay. For six contacts, about one-half of the parts shown will be needed for the U relay, whereas the new relay will again require two wire spring combs. In the new relay three wire spring combs are needed for any contact combination which includes both make and break contacts up to twelve makes and twelve breaks. Four wire spring combs are needed for a relay having twenty-four make contacts.

Two problems had to be solved in providing molded wire spring combs, namely, wire straightening and molding of a multiplicity of wires. Both of these were studied cooperatively at Bell Telephone Laboratories and the Western Electric Company.

Wire is straightened by rotating cam and die members around the unstraightened wire which causes alternating flexing of the wire. For best results, it was found important to shape the cams properly and to

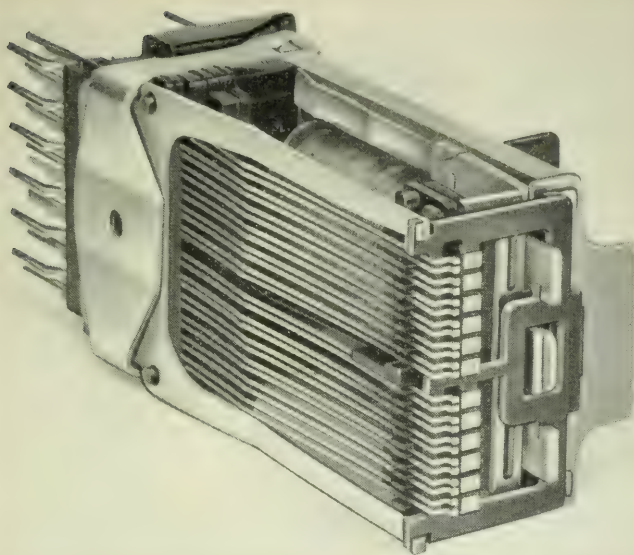


Fig. 10—AF relay with twelve make contacts.

push, rather than to pull, the wire through the rotating cams. By this means it is expected to get straight wire without producing an appreciable twist in it. The Western Electric Company has developed a multiple head wire straightening machine which can be directly associated with the molding press.

A multiplicity of straightened wires is fed into a molding press where plastic molding is used to hold them in proper location. Molding of wire required that the plastic, fed into the die, avoid any appreciable distortion of the wires between unsupported sections. A considerable amount of development work, chiefly by the Western Electric Company engineers, was required to achieve this result. Transfer molding of a thermosetting phenolic plastic has been chosen as the most suitable for producing stable wire spring subassemblies. This is based on the need for stability of the wire positions and because of the ability of the material to withstand the effects of heat. Fig. 12 shows continuous ladders of molded wire spring sections before cutting to length.

The molded sections have a number of features of design importance beyond holding the wires in place. These added features are provided by shaping molded sections to make the remainder of the relay simpler. In particular, these features provide registration pins and holes, guides for

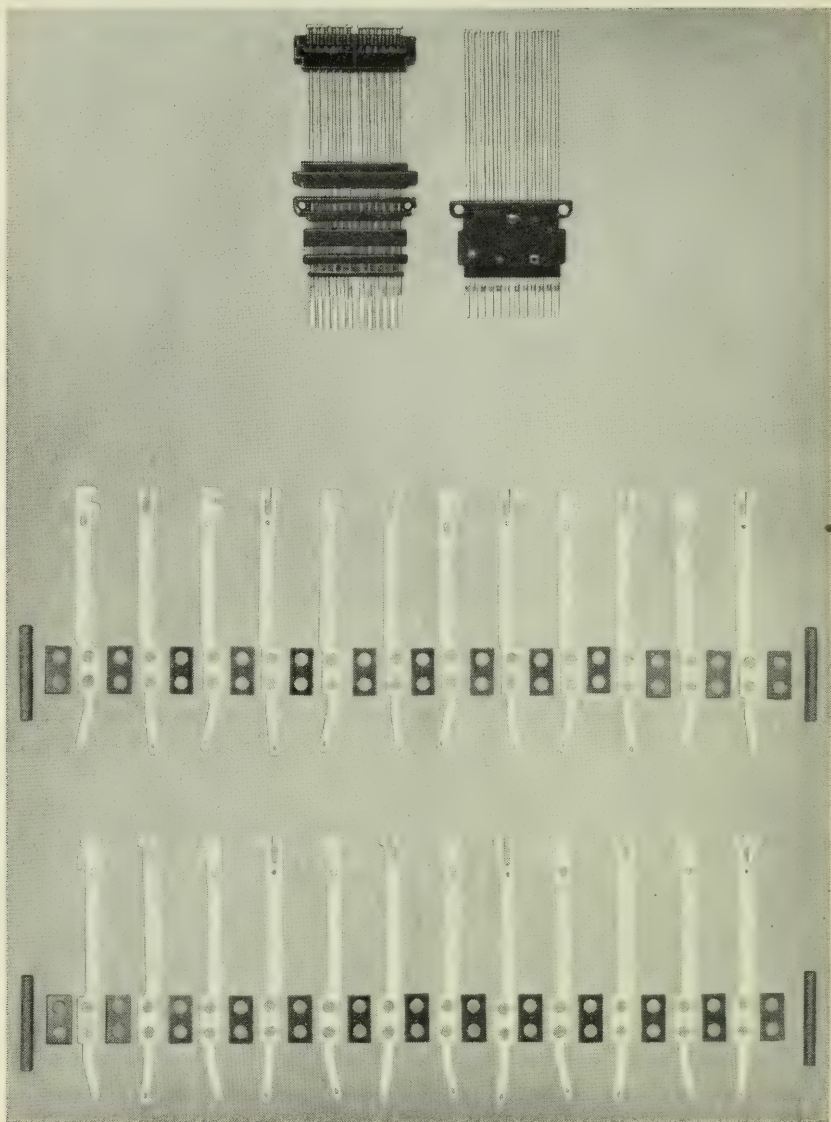


Fig. 11—A comparison of the molded wire assemblies for twelve make contacts with the corresponding U relay parts.

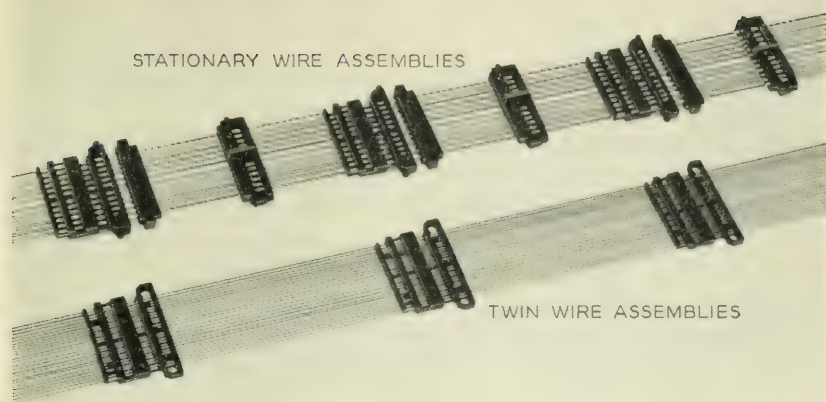


Fig. 12—Molded stationary and twin wire assemblies, before cutting to length.

the ends of the twin wires, cover anchorage, damping material support, etc.

7. CONTACTS AND CONTACT WELDING

Since the primary purpose of the relay is to open and close electrical circuits through the contacts, there has been a special effort to make these contacts as reliable as possible. Accordingly, palladium is used for all contact surfaces. This use of precious metal substantially eliminates opens due to corrosion. Palladium not only gives outstanding reliability but studies indicate that its use results in the best economic balance between manufacturing cost and service because of the reduced maintenance expense.

Open circuits due to particles of dirt between the contact surfaces are largely eliminated by the use of a contact cover, complete independence of the twin contacts described in the section Relay Performance, and the dynamic characteristics of the wire springs. However, to further reduce the incidence of dirt troubles, the surfaces of the twin contacts are coined to a cylindrical shape. This greatly reduces the effective bearing area between the twin contacts and the flat surfaces of the single contacts. Thus, even if an occasional dirt particle should come to rest on one of the contact surfaces, there is small likelihood that it would be in the contact area.

Since welding contacts to wires instead of flat springs is relatively new, considerable attention was given to the development of suitable

techniques. The basic requirements for satisfactory welds are:

1. Sufficient strength to withstand the forces encountered during manufacture and service;
2. Accurate positioning of the contacts on the wires, and
3. Low cost.

These requirements apply to both stationary and twin contacts. However, because of differences in geometry, entirely different methods have been developed for welding the two types of contacts.

The twin contacts are produced by spot welding precious metal contact tapes to the tips of the twin wires. The diagram of the welding circuit is shown in Fig. 13. The condenser *c* is charged by a power supply to a predetermined voltage. The condenser is then discharged through the primary of the welding transformer *T* giving rise to the low voltage

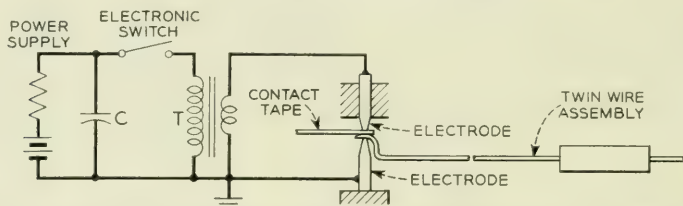


Fig. 13—Diagram showing the essential elements of the spot welding process used for the twin contacts.

high current surge which produces the weld. The contacts are then sheared to length and the surfaces are coined to a cylindrical shape.

The spot welding process did not appear best for welding the stationary contacts to the ends of the wires because of the need to grip the wires with heavy welding electrodes in the limited space directly behind the contacts. Accordingly, a type of welding known as "percussive welding" was developed, which permits one of the electrodes to be placed near the wiring end of the wire springs without developing excessive heat in the wires and which also permits the accurate positioning needed for the contacts in order to control the point of contact closure on the assembled relay. The welding circuit is shown in Fig. 14. The condenser *c* is charged by means of a direct current power supply, and the condenser voltage also appears on the stationary wire. As the contact to be welded is moved toward the end of the wire, the condenser discharges forming an arc which melts the abutting surfaces of the contact and wire. The constants of the electrical circuit and mechanical system were chosen to assure melting a proper amount of metal at a controlled rate to assure high weld strength. The parts are held together during the very brief

cooling period as the weld is completed. A small resistance r is added in series with the discharge circuit to limit the current and control the arcing period.

That high weld strengths are obtained by this process is indicated in Fig. 15 which shows typical distributions of weld strength for both the percussive-welded contacts and the spot-welded twin contacts. The plots show the percent of contacts with weld strengths equal to or less than any prescribed value within the range of the chart. As shown, the percussive welds are generally stronger than the spot welds which is, in part, due to larger welded areas.

Although percussive welding is more suitable for the stationary contacts welded in the factory, it is planned that occasional replacement of both stationary and twin contacts will be made in the field by spot

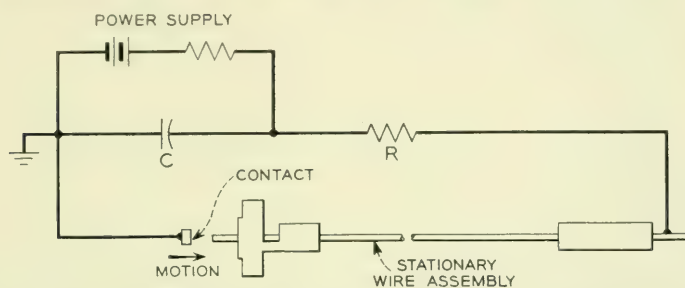


Fig. 14—Diagram showing the essential elements of the percussive welding process used for the stationary contacts.

welding. This will be done with the Bell System field welding equipment⁹ provided with suitable electrodes. In this case, however, more expensive all-palladium stationary contacts of special shape would be used to facilitate the spot welding and individual hand adjustment for final position of the contacts will be necessary.

8. STANDARDIZED ASSEMBLY OF CODED PARTS

Since assembly was one of the most promising fields for reducing costs in a new relay design, special effort was made to reduce the assembly cost of the AF relay. Some of the major design features which contribute to low cost assembly are:

1. The continuous molding and fabricating processes for the wire spring subassemblies, which avoid all individual handling of wires and contacts.

⁹ W. T. Pritchard, "Relay Contact Welder", *Bell Lab. Record*, April, 1944, p. 374.

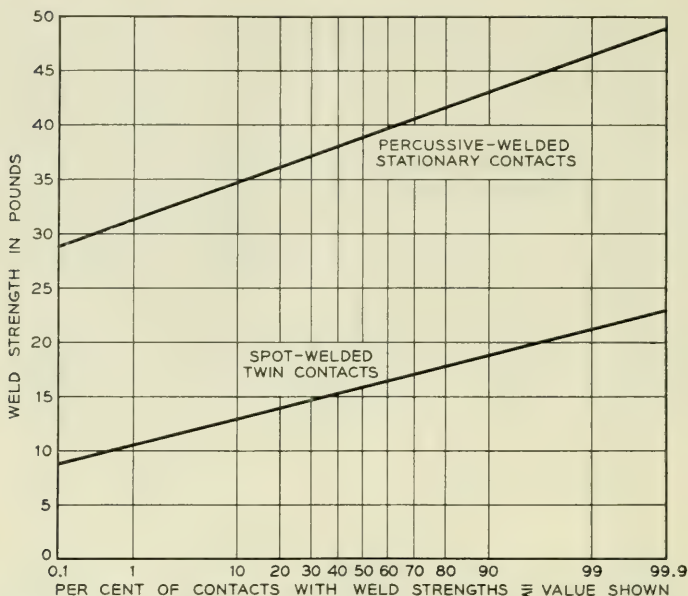


Fig. 15—Typical weld strength distributions for the stationary and twin contacts. The horizontal scale is graduated so that normal distributions will plot as straight lines.

2. Clamping the relay pile-ups by means of a simple spring clamp instead of the more conventional method using screws.

3. A single, easily mounted, operating card.

Less obvious, but equally important is the basic philosophy whereby a large variety of different relay codes are obtained by assembling parts which for assembly purposes are essentially identical for each code. As previously described, the spring combination for each relay is controlled by selection of the proper code card, twin wire assemblies with wires in the proper positions for that combination, and a stationary wire assembly with the right kind of contacts welded to the proper wires. At the present time six different card varieties, fifty twin wire assemblies and seventy-five stationary wire assemblies have been standardized. The twin wire assemblies are provided with any number from one to twelve pairs of wires in various positions while the stationary wire assemblies have from one to twelve contacts in matching positions, with the added variable that each contact may have precious metal on either or both sides as needed. With these it is possible to obtain more than 300 different contact combinations, although only about 100 of these are now needed. Yet, with a few exceptions, each relay code is assembled from

the same number of parts put together in the same manner. By using additional varieties of cards and wire spring assemblies the total number of contact combinations which are possible with the basic design is many times larger than the 300 indicated above.

Other examples of coded parts which are assembled in a standardized manner are the coils, core plates and restoring springs. Although coils vary greatly as to turns, resistance, etc., all are assembled to the cores by the same procedure, using identical spoolheads. The three values of armature travel are controlled by selection of core plates with the proper size of openings, but all core plates are assembled alike. Similarly, the restoring springs are provided in seven varieties including six different thicknesses and seven predeflections to give the desired restoring force, but these variations do not affect the assembly operations.

Standardized assembly of coded parts is of value, not only in reducing the cost of hand assembly operations, but also in providing a more uniform product and as a principle which may make machine assembly practicable.

9. RELAY ADJUSTMENT

Since adjustment expense accounts for a considerable part of the manufacturing costs of older type relays, special efforts were made in the design of the AF relay to reduce the need for adjustment. As a result several types of adjustment used with other relays have been eliminated completely and the remaining adjustments have been simplified. All individual contact adjustment has been eliminated and only two types of factory adjustments are made with the AF relay. These include adjustment of the restoring spring to control armature back tension and a gang adjustment of the stationary contacts to control the points in the armature travel at which the contacts operate. Even these adjustments are needed for only a fraction of the relays as close control of the tolerances in manufacture often causes the back tension and contact operate points to fall within acceptable limits as the relays are assembled.

The gang adjustment of the stationary contacts is made by bending the arms of the core plate, thereby changing the position of the front molded section of the stationary wire assembly which rests on the ends of the arms. Each arm may be bent by means of a screwdriver inserted in the slot as shown in Fig. 16. Rotation of the screwdriver in a counter-clockwise direction causes the upper end of the core plate arm to move to the left, carrying with it the upper end of the stationary wire assembly, including the stationary contacts. This reduces the gap between these

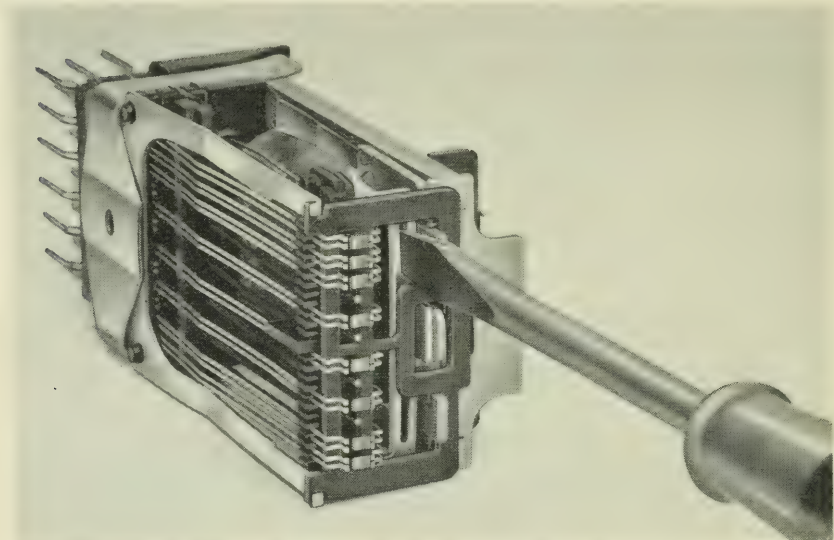


Fig. 16—Contacts may be gang adjusted to operate at the proper points in the armature travel by bending the arms of the core plate with a screwdriver.

stationary contacts and the make twin contacts, thereby causing these contacts to operate earlier in the armature travel. Since the break twin contacts rest against the stationary contacts, these are also moved to the left, reducing the space between the break twin wires and the actuating surface of the card. Thus, bending the core plate arms to the left causes both make and break contacts to operate earlier in the armature travel, while bending the arms to the right causes these contacts to operate later. By bending both arms in the same direction, the operate points of all contacts may be shifted in the same direction. On the other hand, separate arms permit adjustment of the upper relay contacts independently of the lower contacts, thereby increasing the latitude of adjustment.

The parts of the relay are dimensioned so that no adjustment of the core plate arms is required, except to compensate for variations in manufacture of the relay parts. Hence relays assembled from parts made with sufficient accuracy do not generally require adjustment.

The restoring springs may be adjusted for the proper armature back tension by the use of a simple spring bending tool applied to the side arms. However, springs are provided with various predeflections and thicknesses to correspond with various numbers of make twin contacts which must be held open in the unoperated position. Again, no adjustment for back tension is necessary except to compensate for variations

in manufacture, as the restoring spring tension is normally just sufficient to overcome the tension of the make twin wires and hold the armature against the backstop within acceptable force limits. Close control of the tension bends in the wires and restoring springs reduces the frequency with which adjustments are needed and a large portion of the relays do not require this adjustment.

Types of factory adjustment which are common on other relays but which have been eliminated entirely on the AF relay include adjustments for contact force, individual adjustment of contacts for contact operate point, and adjustment for armature travel. Contact force is controlled by means of the large predeflections of the twin wires as mentioned previously. Individual contact adjustment is eliminated by close control of tolerances combined with the single card method of actuation, and by the simpler gang adjustment used when necessary. Adjustment for armature travel is eliminated by the use of close tolerances on the controlling dimensions of the parts.

Adjustments of worn relays in the field may be limited to gang adjustment of the contacts and back tension adjustment as described above. Other adjustments may include burnishing the contacts to remove surface irregularities, replacement of contacts and individual contact adjustment as mentioned previously, and replacement of the card if it should become badly worn or damaged. If card replacement is necessary this may be done without dismounting the relay from the mounting plate and without disconnecting the associated wiring.

10. RELAY PERFORMANCE

As previously stated, the broad objective in the design of the AF relay has been to reduce the annual charges for the use of this relay in the telephone system. Part of this reduction comes from lower manufacturing costs; the remainder comes from savings associated with the improved performance characteristics, such as long life with relatively low maintenance expense, reduced power consumption, and increased speed which reduces the number of units of certain types of equipment, such as markers, needed for telephone central offices. A brief description of some of the principal characteristics of the new relay follows.

Load and Pull Characteristics

Typical load and pull curves for a wire spring relay with twelve early break-make transfer contacts are shown in Fig. 17. The abscissa shows the motion of the armature as it travels from the unoperated position

to the operated position, and back again. This is measured at the center-line of the card and hence is also the card motion. In the unoperated position the armature rests against a backstop, which is part of the core plate. In the operated position it rests against 0.006-inch thick non-magnetic separators which prevent the armature from touching the core. The ordinate shows the spring load, which opposes the armature motion

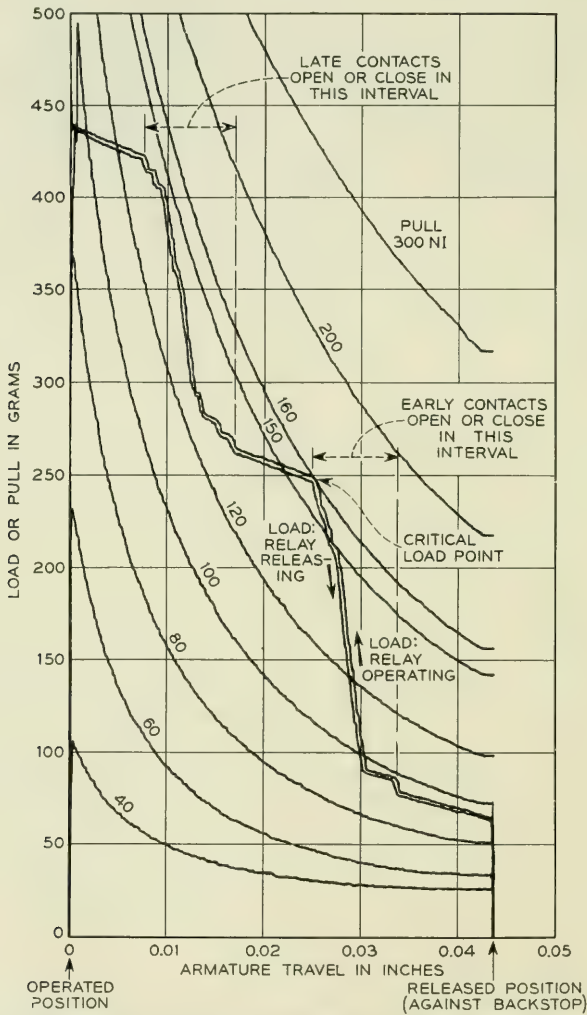


Fig. 17—Typical load and pull characteristics of a wire spring relay with twelve early break-make transfer contacts.

toward the core, and the magnetic pull acting on the armature for various numbers of ampere turns in the winding. These pull and load curves are also measured at the card.

Examination of the load curves shows several features of the relay. The armature back tension, or force, holding the armature against the backstop is about 65 grams in this case. As the armature moves toward the core, the spring load increases along the upper of the two nearly-parallel load curves until it reaches a final value of about 440 grams in the operated position. As the armature is allowed to return to its original position, a second curve, just below the original curve, is obtained. The area between these two curves is a measure of the energy loss due to mechanical hysteresis, or friction, in the relay. As can be seen from the curves, the friction in the new relay is very low and is a small fraction of the spring load at all values of armature travel.

The shape of the load curves is characteristic of AF relays with intermediate travel (0.044 inch). The load increases rapidly in two regions, corresponding to the intervals in which the early and late contacts operate. The rapid increases are caused by the armature and card picking up the additional load of the twin wire springs. Each of the 48 twin wires is picked up almost abruptly at various points and the summation of these additions to the load gives the irregular appearance shown.

The pull curves of Fig. 17 are for essentially static conditions since the armature was restrained to move slowly through its travel while the curves were automatically recorded. These curves are of interest because they show the ampere turns necessary to assure operation of the relay and also values which will assure the armature will not leave the backstop. For example, the "critical load point," or point on the load curve which requires the greatest number of ampere turns, is seen to occur at 0.025-inch travel and 250 grams, which under static conditions would require at least 160 ampere turns in the winding to assure complete operation. On the other hand, as little as 94 ampere turns could cause the armature to leave the backstop and might cause operation of one or two contacts. Hence, a lower value must be maintained to assure that the armature will remain at rest against the backstop. This information is important for relays having non-operate requirements. Similar information may be obtained for limiting ampere turn values which will assure that the armature will remain in the operated position (hold requirements) and, again, which will assure complete release to the backstop position (release requirements).

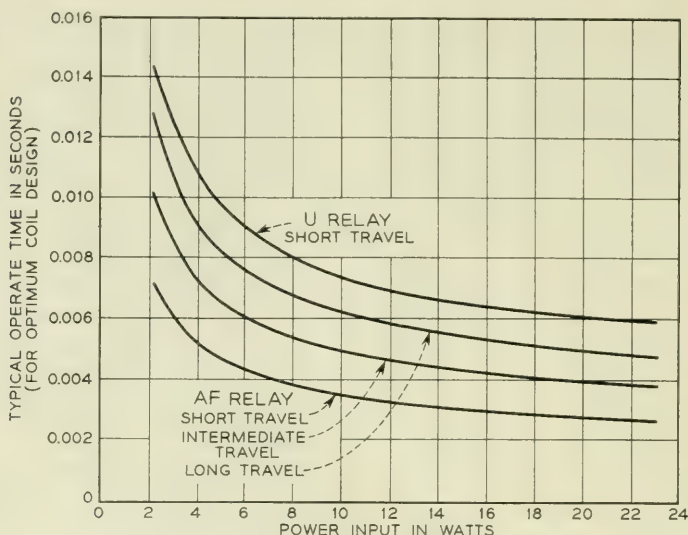


Fig. 18—Typical operate times of speed relays with optimum coil designs for high speed operation.

Speed of Operation and Release

Typical operate and release times of the AF relay are shown in Figs. 18 and 19. Fig. 18 shows the operate times for "speed relays" in which the speed is limited primarily by the time needed to accelerate the mass of the moving system, and is not affected appreciably by the spring loads. These are relays with coil windings of about 1000 ohms resistance, or less, corresponding to power inputs of 2.3 watts, or greater, when connected to a 48-volt supply. For each value of resistance, the operate time shown is obtained with windings having the optimum number of turns for shortest operate time. This time is plotted as a function of power input for various armature travels. Short travel relays have typical operate times varying from about 2.5 to 7 milliseconds as the power is reduced from 23 to 2.3 watts, or as the resistance is increased from 100 to 1000 ohms, using the appropriate number of turns in each case. Intermediate and long travel relays have longer times. For comparison, a short travel U relay is also shown. This relay requires about twice the time to operate as the corresponding AF relay. The improvement with the AF relay is due primarily to the lighter mass of the moving system and slightly shorter travel due to better control of tolerances. Better control is inherent with the single card system for contact actuation and is accomplished without individual adjustment of contact or backstop position, both of which are hand adjusted on U relays.

Typical release times for the AF and U relays are shown in Fig. 19, with time plotted as a function of the number of contact pairs for relays equipped with standard 0.006-inch thick nonmagnetic separators. In this case the improvement is greater than two to one, due principally to the lighter moving parts of the AF relay and lower eddy current effects of the rectangular silicon iron core.

Power Requirements

The nominal power required to assure operation, with some margin, of relays with windings designed for minimum power consumption is shown in Fig. 20. Included is an allowance for adverse variations in magnetic structure, winding and loads. Since least power will be used by the largest coil wound with the finest wire consistent with meeting the ampere-turn requirements for the various contact loads, the curves are discontinuous and have steps as the wire sizes are shifted from one size to the next to meet the ampere turns needed for increasing loads. Again, the corresponding U relay power is shown for comparison. For corresponding numbers of contact pairs, the AF relay requires about half the power of the U relay, except with fewer contact pairs where the power in each case depends upon the use of No. 41 gauge wire. This comparison applies only when coils of optimum design for minimum power are used on both relays. In practice, the coils are selected for the best economic balance between power consumption, cost of the coils and value of speed of operation. Coils designed for minimum power are rela-

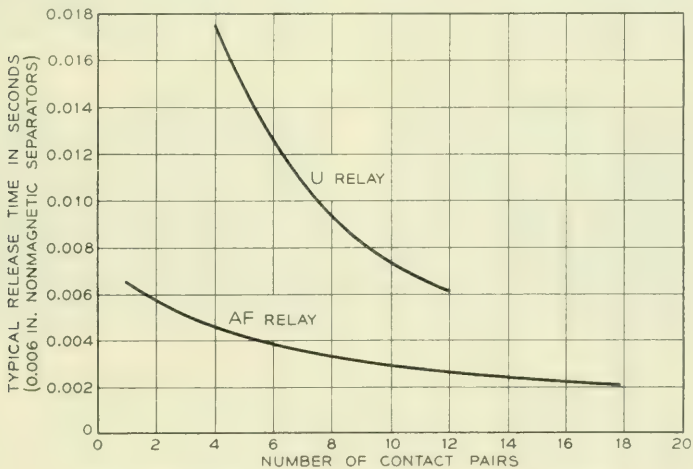


Fig. 19—Typical release times of AF and U relays.

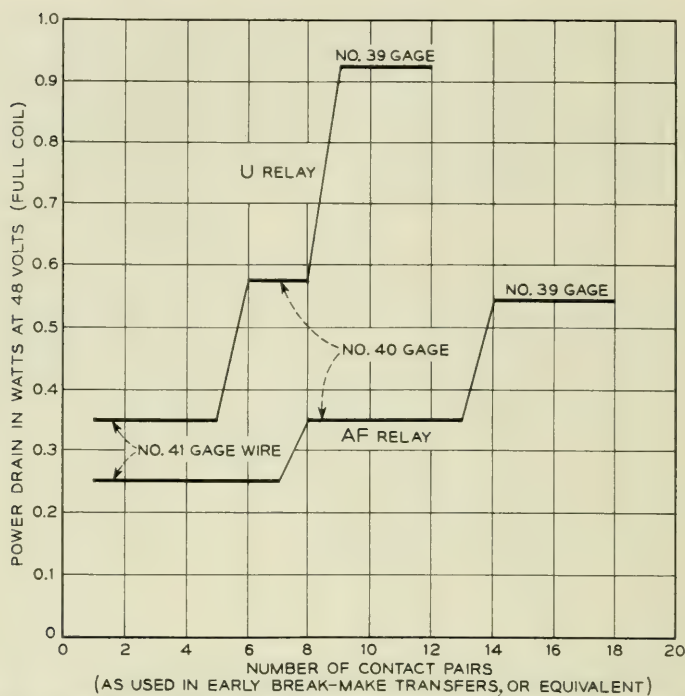


Fig. 20—Power used for AF and U relays with coils designed for least power.

tively expensive because they contain as many turns of fine wire as the available space permits, and their use is economical only on relays which are operated an appreciable portion of the time and where speed is relatively unimportant. The reduced power required for the AF relay is due principally to an improved magnetic structure, shorter travels for similar contact combinations, and lower contact forces.

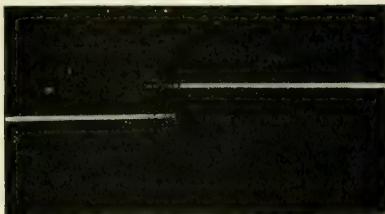
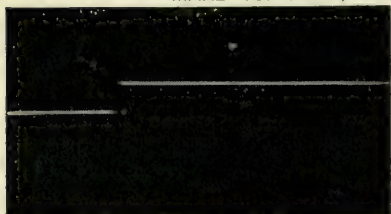
Contact Performance

The principal characteristics which must be considered in evaluating contact performance include chatter, erosion or wear, susceptibility to open contacts and locking, and changes in these characteristics with wear of the relays. In general, all these features are improved on the AF relay compared with the U type.

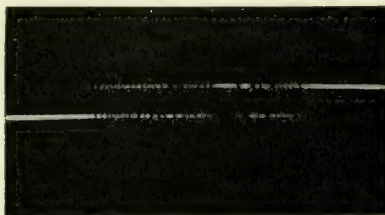
Typical chatter on closure of make and break contacts on U and AF relays built for moderate and fast operation is shown in Fig. 21. The degree of improvement of the AF relay is striking. The reduction in chatter has direct circuit advantages in reducing the possibility of false

AF RELAYSU RELAYS

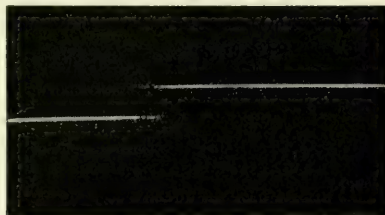
MAKE CONTACTS, MODERATE SPEED OPERATION



MAKE CONTACTS, HIGH SPEED OPERATION



BREAK CONTACTS, MODERATE SPEED RELEASE



BREAK CONTACTS, HIGH SPEED RELEASE

0 1 2 3 4 5 0 1 2 3 4 5
TIME IN MILLISECONDS

Fig. 21—Typical chatter on closure of contacts.

operation of associated high speed equipment and also indirect advantages in prolonging the life of the contacts. The improvement is due largely to the type of card operation of the completely independent, low-mass twin wires and also to the low mass of the moving system which excites less vibration of the relay structure as a whole. The placement of the card close to the contacts allows the full contact force to be developed within a very short time, and the low mass of the twin wires stores little kinetic energy to cause reopening due to wire vibration.

A particularly troublesome type of chatter occasionally experienced is caused by rebound of the armature after striking the backstop. This chatter is objectionable because of its long duration which is of the order of a millisecond and may occur several milliseconds after the initial opening or closure of the contacts. This increases the possibility of false operation of associated circuits. Accordingly, a fundamental study was made of the means for reducing armature rebound, as previously mentioned. As a result, changes were made in the suspension of the armature and in the position of the backstop which substantially eliminated chatter due to armature rebound.

Electrical erosion of the contacts is reduced on the AF relay because of less chatter and because of the lower energy levels controlled by the contacts, where these are used to operate other AF relays. This improved performance not only reduces maintenance but permits the use of less expensive, smaller size contacts.

Contact locking is substantially eliminated on the AF relay because of the card operation, where the static and dynamic forces associated with the card and armature are available to break loose any incipient lock. Open contacts are reduced by (1) protecting the contacts from dust with a small cover, (2) rounding the twin contact surfaces to reduce the effective areas on which particles must lodge to cause opens and to increase the pressure on the areas, (3) the use of palladium contacts, (4) the dynamic characteristics of the wire springs, and (5) the use of twin contacts on completely independent twin wires. The complete separation of the twin wires is an important feature in reducing open contacts. As shown in Fig. 22, the flat punched springs of the U relay carry twin contacts but these are mounted on tips which are separated by a relatively short punched cutout. This limited separation did not achieve the full advantage of twin contacts as a sufficiently large particle of dust under one contact could cause both contacts to be held open. A subsequent design, known as the UB relay,³ used a longer cutout, resulting in greater independence of the twin contacts with a significant reduction in contact opens. The AF wire spring relay achieves complete

separation by use of separate twin wires and a significant part of the improvement with respect to contact opens is due to this feature.

The improvements in contact chatter and open contacts become even more evident during the life of the relay. As shown in Fig. 23, the contact forces on U relays diminish rapidly with wear of the contacts resulting in increased chatter, more frequent opens and the need for earlier readjustment. Card operation of wire springs with large predeflections, however, assures substantially constant forces, thereby maintaining the initial chatter-free performance and fewer open contacts.

Life

Tests of relays with contacts protected electrically with resistance-condenser networks indicate that the standard AF relays with winding resistances of 700 ohms or greater will have a life in the order of 250–500 million operations before readjustment becomes necessary. With readjustment, of course, these figures can be increased several times.

Where longer life is essential, special features are used to increase the life. With these features a life, before readjustment, of a billion operations is expected for some relays and, with readjustment, all relays equipped with the special features for long life should be capable of a billion operations.

The special features for long life include vibration dampers attached to the twin wires and the stationary wire assembly as shown in Fig. 24, wear-resistant nickel and chromium plate on the armature, core, and core plate, and a long-wearing alloy for the nonmagnetic separators

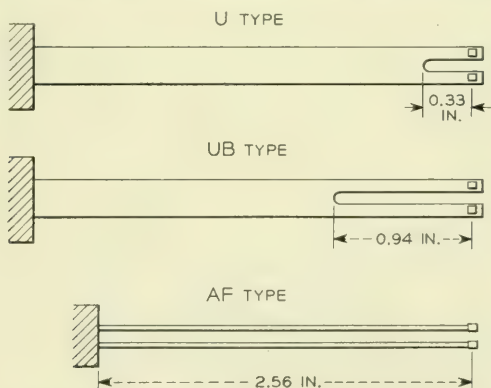


Fig. 22—Independent action of the twin contacts is limited on U and UB relays because both contacts are mounted on a single spring which is notched. The AF relay achieves complete independence by mounting the contacts on separate wires.

welded to the armature. The special features consist largely of variations in finish and material which do not greatly affect the manufacturing processes. The only added parts are the damping members. These are molded from soft but stable polyisobutylene with grooves to receive the twin wires. One damper is attached to each side of the shelf provided on the stationary wire assembly. The twin wires pass through the grooves and are cemented in place. As shown in Fig. 25, these dampers reduce the vibration of the twin wires between the card and the molded section at the rear, thereby reducing the slide between the wires and the card.

Early designs of relays indicated that wear between the twin wires and the card was excessive and that changes in materials would not produce the improvement needed for very long life, particularly with high-speed relays. A fundamental study¹⁰ of the conditions which cause wear was made and it was found that reduction of the sliding motion between the wires and card to 0.001 inch or less was necessary to substantially eliminate such wear. The AF relay meets this requirement. The necessity for such a requirement will be better understood when it is

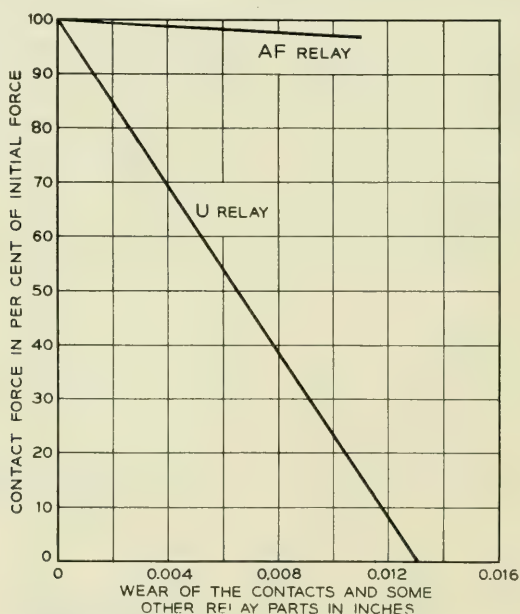


Fig. 23—Contact forces on the AF relay remain almost constant with wear, while U relay contacts lose force rapidly.

¹⁰ W. P. Mason and S. D. White, "New Techniques for Measuring Forces and Wear in Telephone Switching Apparatus", *Bell System Tech. J.*, May, 1952, p. 469.

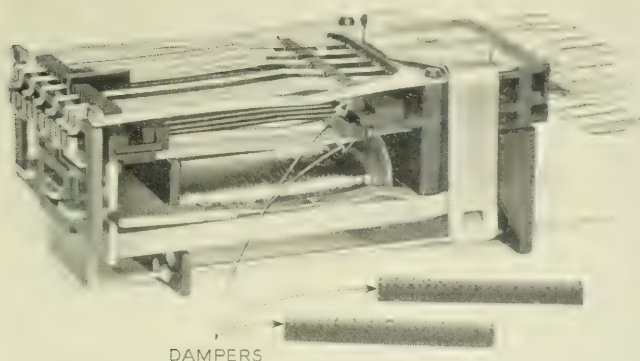


Fig. 24—Where very long life is needed, polyisobutylene dampers are mounted between the twin wires and a molded shelf on the stationary wire assembly.

noted that, for one billion operations, the total slide corresponds to a distance of about thirty-two miles.

Stability

The AF relay is a distinct improvement in stability compared with earlier designs when subjected to shock or temperature and humidity changes. Under severe and repeated variations in temperature and humidity, the largest changes in contact separation are not more than 0.002

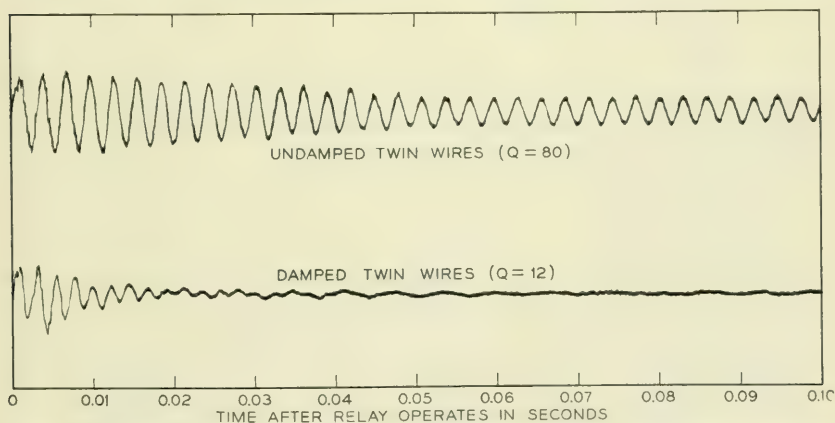


Fig. 25—Oscillograms showing the effectiveness of the polyisobutylene dampers in reducing the vibration of twin wires following operation of the relay. The vibration is measured in the horizontal plane, about midway along the length of the wires.

to 0.003 inch. The improved stability is expected to permit final adjustment and inspection of the relay at the time it is assembled without need for readjustment after it is wired into equipment and installed into service after shipment.

Magnetic Interference

In the past it has often been necessary to maintain large spacing between relays where critical values of current to operate or release the relays must be maintained. In some cases special iron shields were used for further magnetic isolation. Without these precautions, the leakage flux from adjacent relays entered the magnetic circuit of the critical relays and the operate or release currents varied according to whether the adjacent relays were energized.

Magnetic interference between AF relays is substantially eliminated as shown in Fig. 26. This is largely because of the low reluctance of the magnetic circuit resulting from the one-piece core and the large pole face areas between the core and armature. As shown in the figure, the

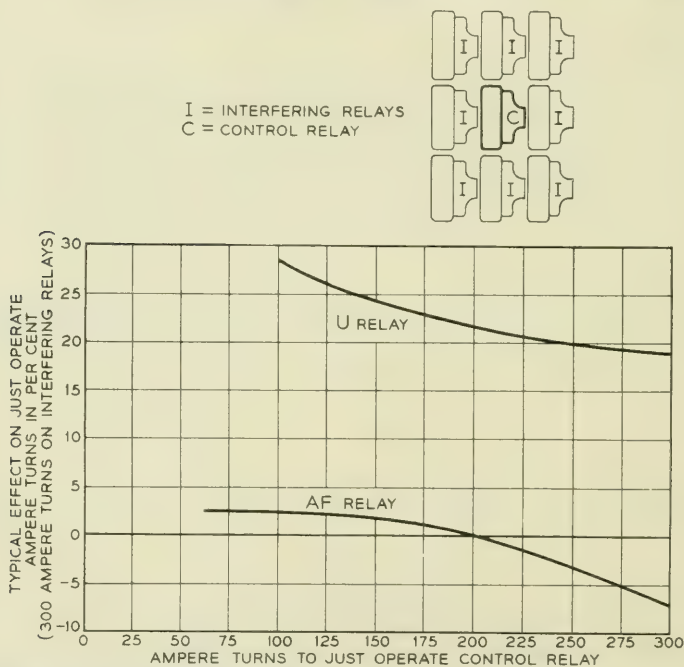


Fig. 26—Typical magnetic interference between AF relays and between U relays, with the relays mounted in the pattern shown.

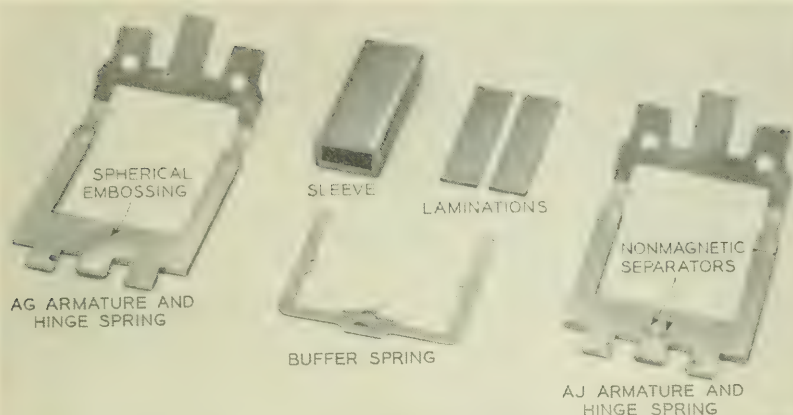


Fig. 27—Additional parts for AG and AJ relays.

measurements were made by surrounding a control relay with eight adjacent closely-spaced interfering relays. The ampere turns to just operate the control relay were varied by changing the mechanical load on the relay, and for each value the change caused by simultaneously energizing the adjacent relays was observed. The improvement of AF relays with respect to the U type is seen to be of the order of ten times for most of the range, with the effects of the adjacent relays being well under 10 per cent up to 300 ampere turns. This is small enough so that no shields or precautions with respect to spacing are required.

11. AG AND AJ TYPE RELAYS

The AG and AJ type relays include modifications of the basic AF design to provide slow release, sensitive, marginal and other additional characteristics. For the most part these modifications are not extensive and the assembled relays closely resemble the AF design.

The additional parts most often used in the AG and AJ relays are shown in Fig. 27. Both relays use thicker armatures with longer side legs than the AF relay, and the armature of the AG relay has a spherical embossing instead of nonmagnetic separators. This reduces the magnetic circuit reluctance of the AG relay when it is in the operated position. In addition, for longer release times, a metal sleeve is assembled over the middle leg of the core, inside the coil. Induced eddy currents in this sleeve oppose rapid changes of flux through the core.

The use of the heavy, embossed armature and a sleeve are sufficient

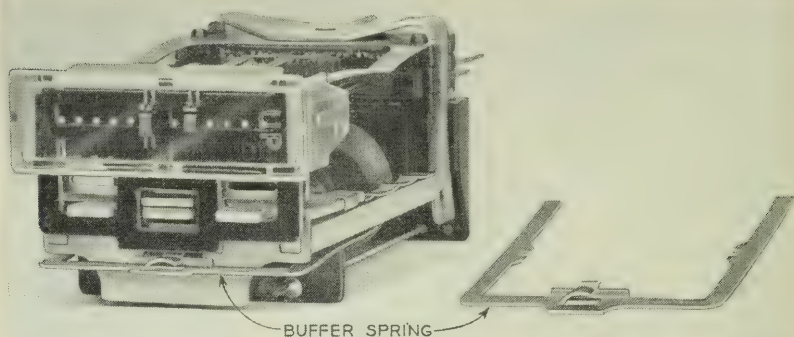


Fig. 28—The buffer spring is used to control the operated spring load, and therefore the release current and release time.

to make the relay slow to release.¹¹ When the current in the winding of such a relay is interrupted, the flux decays slowly due to the circulating currents in the sleeve. Also, the low magnetic reluctance increases the time for the flux decay by permitting relatively high flux values to be maintained by smaller circulating currents. However, to achieve better control of the release times and to maintain stable adjustment during the life of the relays, the following additional features are used:

1. The cores are annealed in a hydrogen atmosphere, chiefly to stabilize the coercive force of the iron.
2. The core and armature have a wear-resisting chromium plate finish to maintain the nonmagnetic gap between the embossed surface of the armature and the core.
3. The use of a spherical embossing reduces variations in reluctance caused by small angular misalignments between the armature and core.
4. Four sleeves are available including light, medium and heavy copper sleeves and a light aluminum sleeve. These sleeves provide various ranges of release time.
5. A buffer spring is provided on the relay to control the operated load and therefore the release time. As shown in Fig. 28, the buffer spring is normally tensioned against the end of the middle core leg. As the relay operates, however, the card strikes the adjustable tab in front of the middle leg and lifts the spring away from the core so that the spring tension is added to the operated load of the relay. The spring may be adjusted for any desired tension, within limits, and the tab can

¹¹ H. N. Wagar, "Slow Acting Relays", *Bell Lab. Record*, April, 1948, p. 161.

be adjusted so that the load may be picked up at any desired point in the armature travel.

When a relay is designed for a specified release time, the spread between maximum and minimum times obtained with a particular sleeve is usually greater than desired. Accordingly a sleeve is selected which under normal conditions would produce a somewhat longer time than the specified value. This time is then reduced, as needed, by increasing the buffer spring tension.

Typical release times plotted as a function of the contact load for AG relays with and without sleeves are shown in Fig. 29. Since this figure illustrates release times that are characteristic of the various sleeves, no buffer spring tension is assumed. As would be expected, the heavier sleeves produce the longer release times, which are also greater for relays with fewer contacts. Even the light aluminum sleeve produces several times longer release times than no sleeve at all. For comparison, release times are also shown for the AJ relay, which has a magnetic structure similar to the AG but with nonmagnetic separators in place of the spherical embossing on the armature. The difference between the AJ

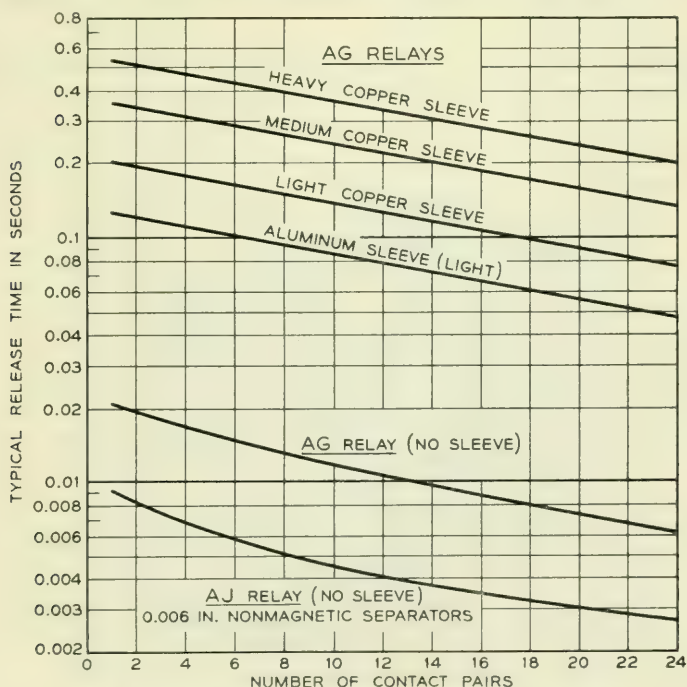


Fig. 29—Typical release times of AG and AJ relays.

release times and those for the AG relay with no sleeve shows the time advantage obtained with the domed armature.

The AJ relay, with its long and relatively heavy armature, is suited for the more critical marginal applications and is capable of operating heavier contact loads than the AF relay. All relays with more than eighteen contact pairs are provided only on the AJ structure. For example, Fig. 30 shows an AJ relay with a full complement of 12 transfers.

A measure of the power requirements for the AJ relay is given in Fig. 31. This shows the power required to assure operation with various numbers of contact pairs for coils designed to consume minimum power at 48 volts. The chart includes allowances for variations in load, magnetic structure and coil, with some margin for changes in these characteristics. Comparison with Fig. 20 shows the power requirements to be slightly lower than for the AF relay except for small numbers of contacts where limitation of wire sizes of the coils is controlling. Under limiting conditions the AJ relay will operate on as little as 0.025 watt.

Other features which may be used to extend the use of the AJ relay for special marginal applications include armatures with various thicknesses of nonmagnetic separators, wire spring assemblies with reduced contact forces, core laminations (strips of iron placed inside the coil, against the middle leg of the core to increase the effective cross-section) and the buffer spring which may be used to control the operated load of the relay and therefore the hold and release currents.

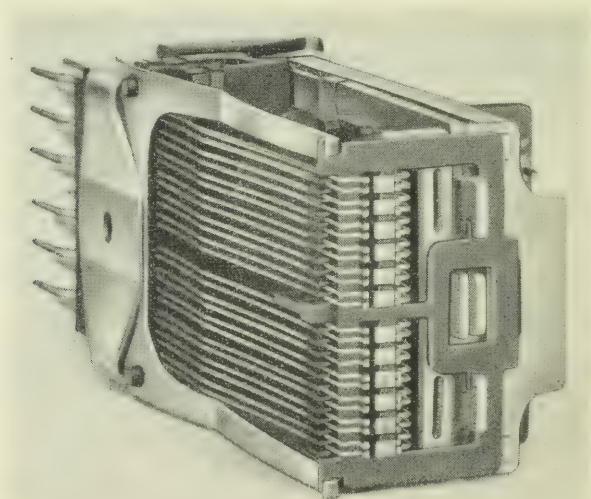


Fig. 30—AJ relay with twelve transfer contacts.

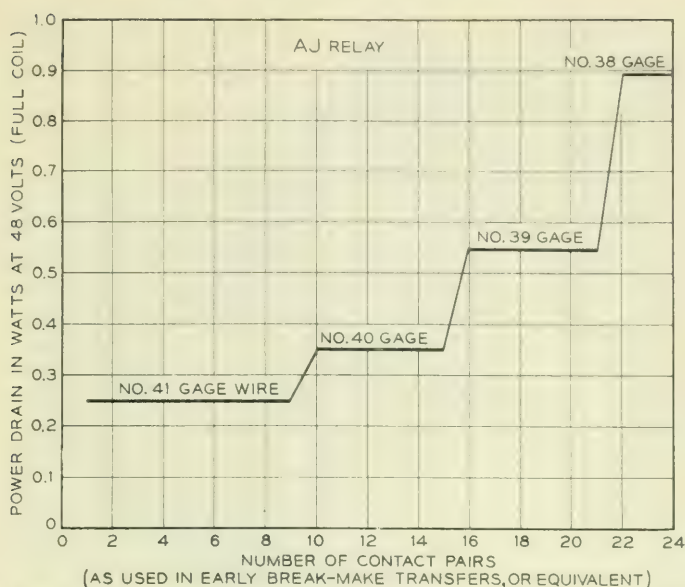


Fig. 31—Power used for AJ relays with coils designed for least power.

A special variety of AJ relay is provided with twenty-four pairs of make contacts as shown in Fig. 32. This relay uses four layers of wire springs and a number of other special parts. As a result, it is often possible to use one twenty-four make contact AJ relay rather than two relays with fewer contacts on each.

12. WIRING THE RELAYS

Connecting wires to the wire spring relay terminals presented a problem which was solved not only for the new relay but by methods which have become important and useful for other apparatus also. The solution came from the invention of a tool, first proposed by H. A. Miloche,¹² for quickly and easily wrapping the connecting wire to the straight end of the wire spring relay terminal.

Early suggestions for making connections to the wire springs included various hooks or bends to simulate the common flat punched terminal having either a hole or notch to facilitate attaching the wires. All of these added some expense to the manufacture of the relay. The added costs were due to two factors (1) forming the wire spring ends required

¹² H. A. Miloche, "Mechanically Wrapped Connections", *Bell Lab. Record*, July, 1951, p. 307.

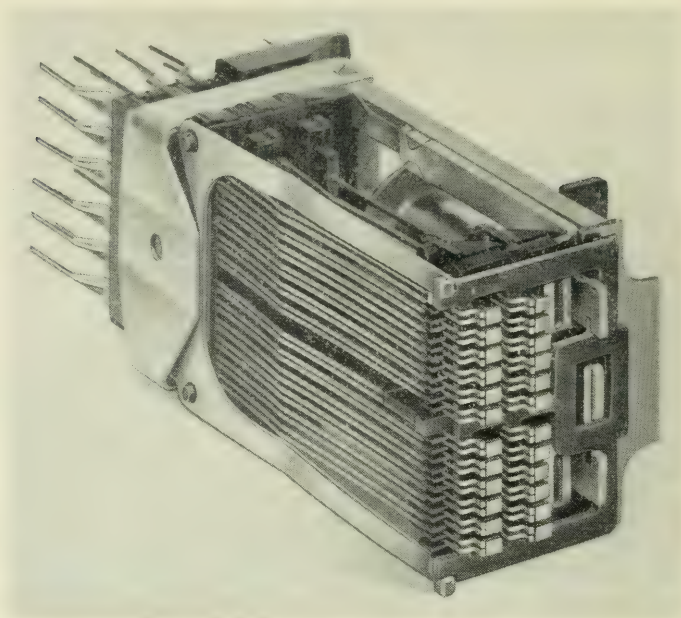


Fig. 32—AJ relay with twenty-four make contacts.

additional operations and (2) the greater flexibility of the wire spring terminals caused some difficulty for the operator so that some increase resulted in the time required to make a connection.

The wrapping tool was first visualized as a simple, trigger-operated hand tool, later as an air or electrically operated tool and still later as a combination tool to do additional operations including cutting and skinning the connecting wire.

Although the wrapping tool was developed to solve a problem which arose in the development of the wire spring relay, it was first applied in commercial practice by the Western Electric Company for wiring to the flat spring relay terminals. Wrapped connections are now used extensively with these terminals, resulting in an improved product at a lower cost. In making wrapped connections to either flat or round terminals, it was the expectation that tinned terminals and wire would be used and soldered together to give a stable, low resistance junction. More recently, however, it has been possible to show that soldering is not required if certain definite dimensional conditions are satisfied by the terminal and the wrapping tool. Solderless wrapped connections will be described in detail in a separate paper.

Fig. 33—Muzzle of wiring tool for making wrapped connections.

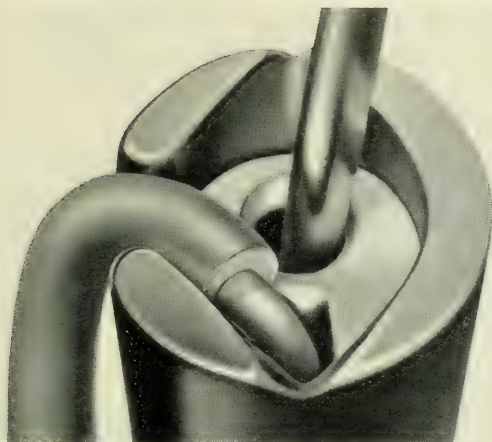
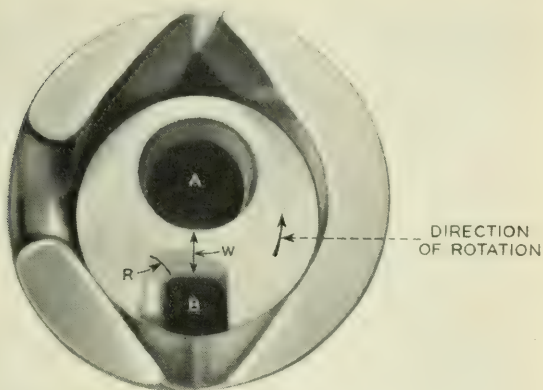
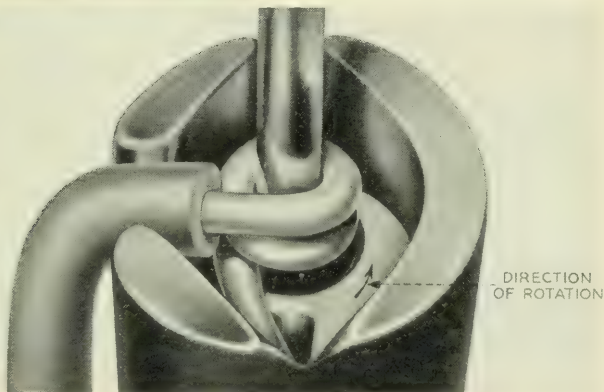


Fig. 34—Muzzle of wiring tool showing terminal and connecting wire in position ready for wrapping.

Fig. 35—Muzzle of wiring tool showing two turns wrapped by rotation of the inner cylinder.



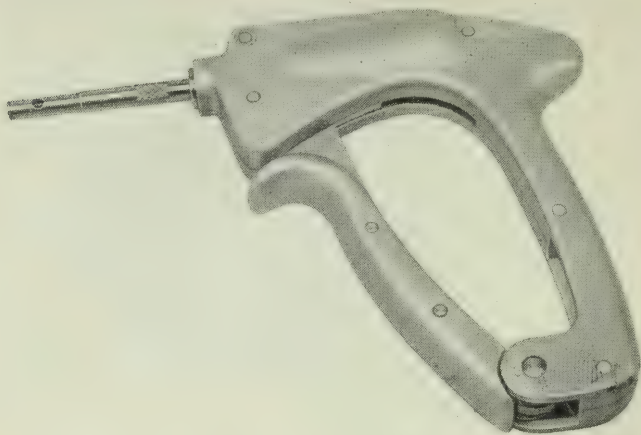


Fig. 36—Hand operated wrapping tool for installation and repair service.

Basic Principles of Wrapping Tools

A drawing of the arrangement and action of a wire wrapping tool is shown in Fig. 33. There are a number of dimensions of the tool and of the terminal which have engineering importance. For the purposes of this paper, it will suffice to note that the radius, r , and the wall thickness, w , are important in producing the best wrapped connection.

As shown in Figs. 34 and 35, the inner member of the tool rotates

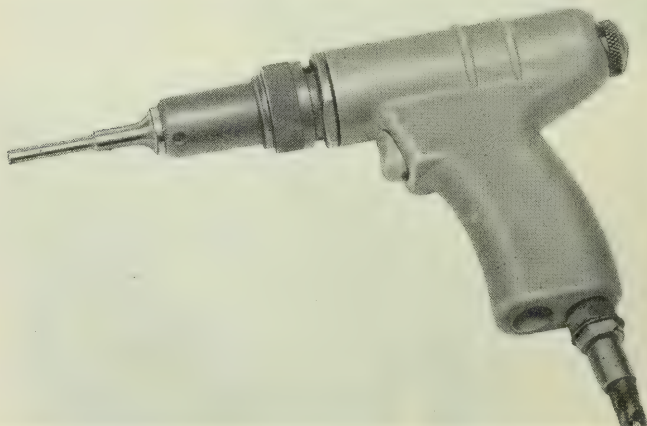


Fig. 37—Wrapping tool operated by air pressure for factory use.

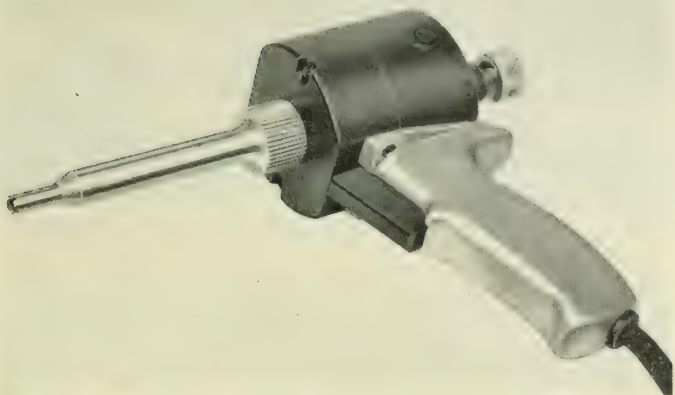


Fig. 38—Wrapping tool operated by electric motor for factory use.

around the terminal which is inserted in hole A of the rotating member. The connecting wire is inserted at B and is anchored by a slight force against the outer stationary member of the tool. As the inner member rotates, the connecting wire is stretched and formed around the terminal until all of the wire length is used. It should be noted particularly that all of the wire is used, making it unnecessary to clip a wire end as in other wiring methods. This is an important detail in avoiding wire clippings which sometimes cause unwanted cross connections in wired equipment units. The tool tip described can be operated by a hand trigger or by motor. Fig. 36 shows a hand powered tool primarily for installation and repair service. Fig. 37 shows a production type tool driven by air pressure developed by the Western Electric Company at the Hawthorne Works. Fig. 38 shows a tool driven by an electric motor used by the Kearny Works of the Western Electric Company. Fig. 39 shows a wrapped connection on a wire spring relay prior to soldering.

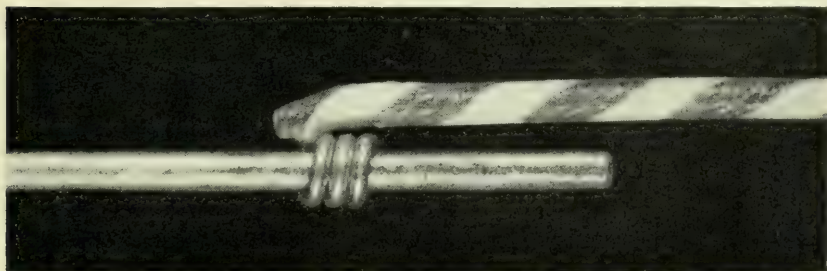


Fig. 39—A wrapped connection on a wire spring terminal, before soldering.

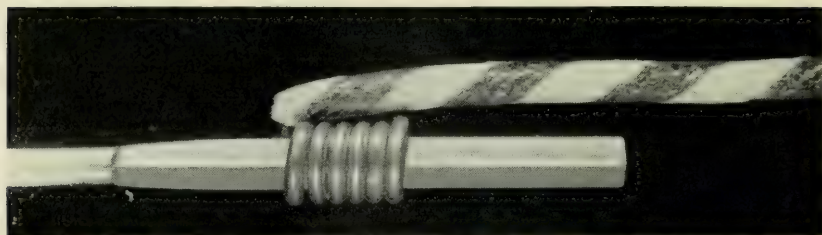


Fig. 40—A solderless wrapped connection on a wire spring terminal.

Another wrapped connection of the solderless type on a similar terminal is shown in Fig. 40. Studies indicate that solderless wrapped connections can be used with a wide variety of materials, including aluminum.

It is interesting to note that a troublesome problem in wiring to the new wire spring relay was solved by the development of a new method which itself has become an important development with broad applications. The wrapped connection with or without subsequent soldering has resulted in better, more uniform and less costly connections made in less time than those made by previous methods.

13. CONCLUSIONS

A description has been given of a new type wire spring general purpose relay for telephone switching systems. Although accurate manufacturing costs will not be available for some time, the new relay is expected to be substantially lower in cost. It provides major improvements in contact performance, reduced power, faster operation and longer life. The new relay also covers a wider field of application than any previous general purpose relay in such characteristics as speed, slow release, marginal operation, number of contacts, etc.

Important economic advantages include lower manufacturing cost of the relay itself and a reduction in switching systems costs resulting from less equipment and reduced power.

The new relay has shown considerable improvement in mechanical life and in contact performance. The life of the relay is of the order of one billion operations. These improvements can be expected to reduce the cost of maintenance of switching systems appreciably.

The development of the new relay has called for a major cooperative effort of many sections of Bell Telephone Laboratories, including such departments as Switching Apparatus Development, Switching Systems Engineering, Switching Systems Development, Research, Materials, Chemical, etc. Without the cooperation of the many members of these

organizations and their special skills, this development would not have resulted in the important and balanced design which has been described. Throughout the development, the associated organizations in the Western Electric Company and the American Telephone and Telegraph Company have made important and guiding contributions, and the New York Telephone Company has cooperated in field trials.

As the spokesman for the project, I wish to express appreciation to the many people who have contributed to this important technical accomplishment. The few names which have been mentioned in the paper were given for historical reasons and cannot replace the large number of important contributors to the project.

To D. C. Koehler, I am indebted for assistance in preparing the paper and the illustrations.

Comparison of Mobile Radio Transmission at 150, 450, 900, and 3700 Mc

By W. RAE YOUNG, JR.

(Manuscript received August 22, 1952)

Based on a series of experiments, a comparison is made of the transmission performance of 150, 450, 900, and 3700 mc in a mobile radiotelephone type of service. This comparison indicates that 450 mc is superior transmission-wise to the presently used 150-mc band in urban and suburban areas. In fact a broad optimum in performance falls in the neighborhood of 500 mc. It is concluded that this range of frequencies would be well suited for providing coverage to meet the large scale needs which are anticipated in and around metropolitan areas. Although higher frequencies are less desirable, the tests indicate that 900 mc is somewhat to be favored over 150 mc from a transmission standpoint if full use is made of the possible antenna gain. Above this frequency, transmission performance falls off even assuming the maximum practical antenna gain. Transmission at 3700 mc suffers an additional impairment in that the fluctuations in received carrier level occur at an audible rate as the mobile unit moves at normal speeds. It is concluded that while transmission above roughly 1000 mc for these services is not impossible, it would be decidedly more difficult to employ these frequencies satisfactorily.

INTRODUCTION

From the beginning of mobile radiotelephone services offered by the Telephone Companies, both "general" and "private-line" types, it has been apparent that the number of channel frequencies then allocated for these uses would not be sufficient to meet the service needs in the near future.

The bulk of these needs will be for service in urban and suburban areas, where business activities are concentrated. These areas are now served on a few individual FM channels in the vicinity of 150 mc. However, a larger number of channels, needed to meet anticipated demands and to develop a more efficient system, are not to be found in the

150 mc region. This space is already allocated fully and permanently to a variety of other services. In fact, this situation extends up to about 400 mc. The larger number of channels for these services apparently will have to be found, therefore, above 400 mc.

However, it is essential to know whether these higher frequencies would be suitable for urban mobile telephone service, or whether there exists an upper limit to the suitable frequencies. In order to answer these questions, a series of tests has been made to compare the adequacy of coverage that could be provided at several representative higher frequencies. These tests were conducted in and around New York City. This location is considered to be typical of the larger metropolitan areas.

THE PROBLEM OF EVALUATION

It became apparent early in the tests that it would neither be practical nor accurate to compare service results for the different frequencies by the method of determining the coverage at the various frequencies, and then comparing these. This would have required, among other things, that "coverage" be defined precisely and then measured accurately in order to determine the differences with the desired accuracy.

Instead, it was recognized that commercial coverage is at present considered to extend into areas wherein a small percentage of the locations will have less than commercial grade of transmission. This might be ten per cent, for example. It was further recognized that, while there existed a trend of performance with frequency, comparative tests at any one location showed variations from that trend. Thus, even if transmitter powers were adjusted so as to offset the transmission effects of that trend, performance at any location would not be equal at all frequencies. But while one frequency might give relatively poor transmission in one location, it might give good transmission at another location, etc. Thus, while the locations of poor transmission were found to be different at the various frequencies, the number of such locations would be the same at all frequencies, provided the trend had been offset by adjustment of transmitter power.

Viewing the problem in this way, it was sufficient to test at enough locations in representative territory to establish this trend in a statistical manner.

Other problems in evaluating differences in suitability of different frequencies lay in how to take into account differences in practical antenna gains and differences in frequency stability. These will be discussed in the next sections.

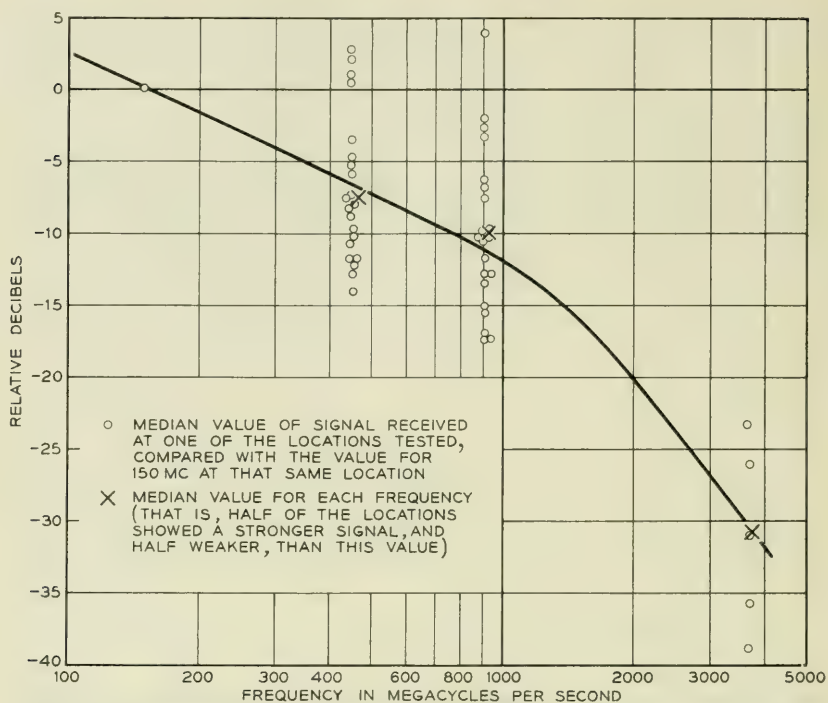


Fig. 1—Median values of received signal power at suburban locations. (Assumes the same power at all frequencies radiated from a dipole and received on a quarter-wave whip.)

OVER-ALL RESULTS

The results of many measurements of path loss between a land radio transmitter and a mobile receiver establish a trend of loss increasing with frequency. This is illustrated in Fig. 1 by the "crosses" which show the strengths of the received signal at higher frequencies as compared with those at 150 mc. The derivation of the values given by the crosses will be discussed in a later section. In the other direction of transmission it appears justified, based upon reciprocal relationships, to assume that path losses from mobile transmitter to land receiver will follow the same trend.

However, although the received signal is seen to decrease with frequency, the amount of received signal which is required to produce satisfactory communication also changes with frequency. The median level of signal required at a mobile or land receiver at various frequencies to override RF noise is given in Fig. 2. The dots here represent the average of many measurements.

Transmitter power required to achieve the same service result at various frequencies has been derived by taking into account the changes of path loss with frequency and also the changes of signal required with frequency. Fig. 3 shows the amounts of power that are required in order to achieve the same coverage in all cases as is now obtained at 150 mc with 250 watts of land transmitter power radiated from a dipole. As shown, the use of an antenna having gain can appreciably lower the land transmitter power that is required. The mobile transmitter power is much less than required of a land transmitter due to the assumption that there are six land receivers located appropriately in the coverage area, rather than just one.

It is apparent from Fig. 3 that the required transmitter power is a minimum in both directions of transmission at around 500 mc. It is also apparent that above this point the required transmitter power increases rapidly with frequency.

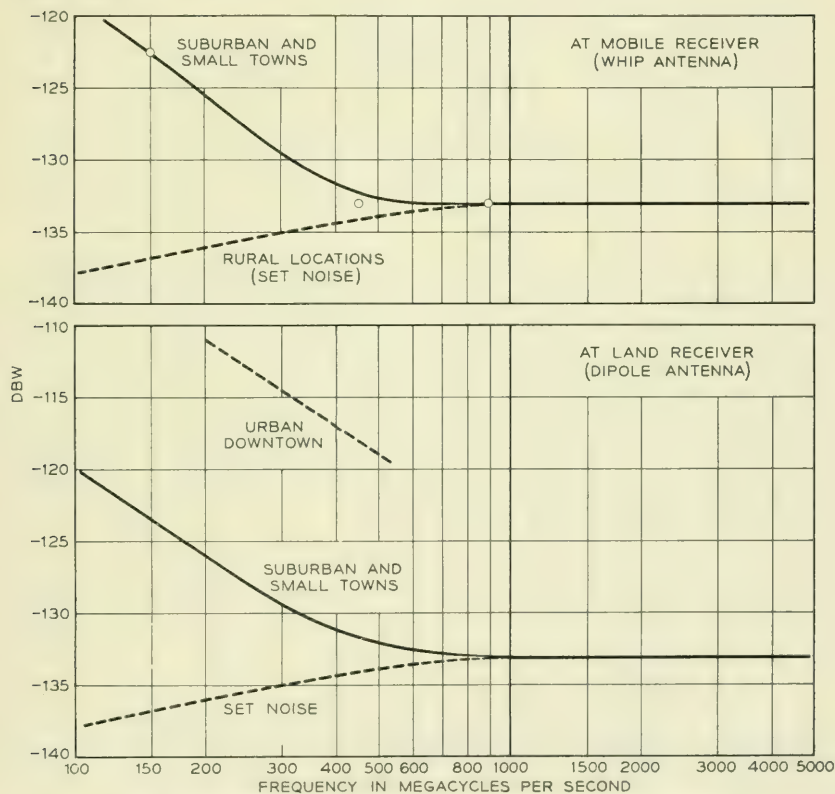


Fig. 2—Median value of signal required to over-ride noise.

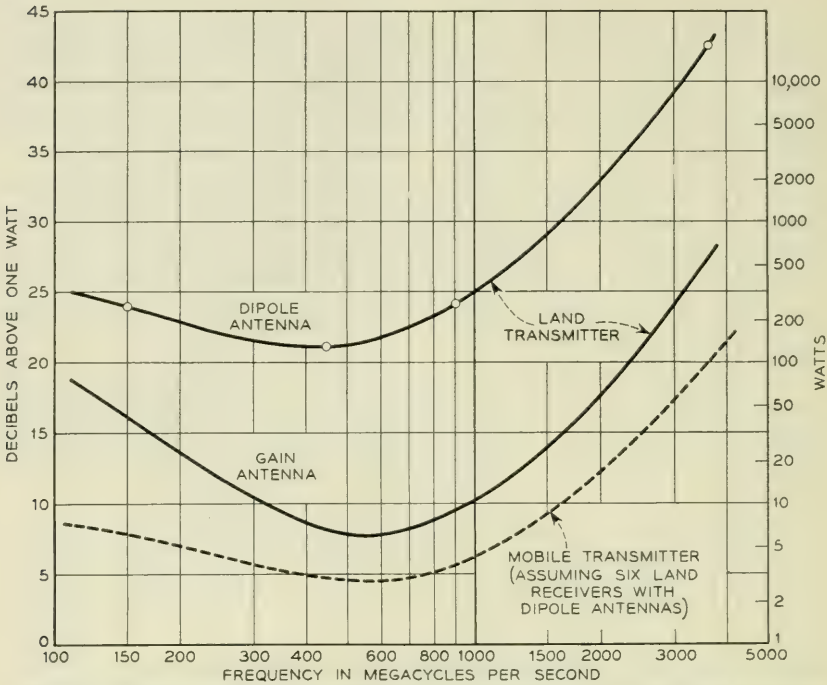


Fig. 3—Transmitter power at antenna input required for urban and suburban coverage. (Mobile antennas are assumed to be quarter-wave whips.)

A word of explanation is needed at this point about the gain antennas which were assumed in one of the curves of Fig. 3. These are antennas which tend to concentrate radiation toward the horizon in all directions. Limits for the amount of gain were based upon the considerations (1) that a set of radiating elements greater than about 50 feet in extent would be impractical to build for this service, and (2) that the vertical width of the beam should not be less than about 2 degrees in order that valleys and hilltops will be covered. The amounts of gain possible within these limits are as follows:

Frequency mc.	Gain-db
150	8
450	13
900	15
3700	15

The mobile antennas were assumed to be quarter-wave whips or the equivalent.

Use of gain antennas for the land receivers would result in still further lowering the required mobile transmitter power. This is not shown on Fig. 3 because the amount of reduction cannot be accurately stated on the basis of present knowledge. It appears certain that the reduction will be at least equal to the antenna gain, and may be appreciably more than this, as indicated later.

The system modulation and pass-band were assumed in the above discussion to be the same at all frequencies. This would not be realistic if the tolerance allowed for frequency instability were a fixed percentage of operating frequency. It may be justified, however, because the necessity for frequency economy and for best transmission performance demands better percentage stability at higher frequencies.

A spot check of transmission, observing circuit merits by listening, has been made to determine the validity of the above results in a very general way. Land transmitter powers were adjusted so that the equivalent dipole power at 450 mc was 3 db less than at 150, and power at 900 mc was 1 db less than at 150 mc. This approximates the powers shown on the "dipole" curve of Fig. 3. The map of Fig. 4 shows the results of this test. While the comparison of circuit merits generally shows a preferred frequency at any given location, the performance appears to be about equal when all locations are considered.

TEST EQUIPMENT ARRANGEMENTS

Tests of transmission outward from the land transmitting station were made on signals radiated from antennas on the roof of the Long Lines Building, 32 Avenue of the Americas, New York City. These antennas were 450 feet above ground. One of the existing Mobile Service transmitters served for the 150-mc tests. Special experimental transmitters were set up for the 450, 900, and 3700-mc tests. All were capable of frequency modulation.

The mobile unit was a station wagon equipped to receive and measure signals at the various frequencies. The receiving equipment was arranged for rapid conversion from 150 to 450 to 900 mc. The bandwidth (about 50 kc) and system modulation (± 10 kc) were identical at all three frequencies (equal to the existing standards at 150 mc). The 3700-mc tests were handled separately. It was not possible to employ the same bandwidth and deviation, but this does not invalidate the comparison of signal propagation at the various frequencies.

A most useful tool in making these measurements was a device known as a "Level Distribution Recorder", or simply "LDR". This was built

especially for these tests and is similar to its forerunners which have been used in the past for measuring atmospheric static noise. The LDR, in combination with a calibrated radio receiver, is capable of taking as many as twenty instantaneous samples of radio signal strength per second, sorting the samples by amplitude, and rendering information on a "batch" of samples from which a statistical distribution curve can be plotted. The LDR was also used for measuring the statistical distribution of audio noise in the output of the radio receiver. The LDR was, in this case, associated with a special converter whose characteristics resemble those of a 2B noise measuring set.

No arrangements were made for measuring radio propagation from mobile unit to a land receiver. It was felt that the comparison by frequencies would be substantially the same as in the outward direction of transmission. It does not follow, however, that the background electrical noise, against which an r-f signal must compete, will be the same at mobile and land receivers. Strength of r-f signal required at land receivers for satisfactory transmission was measured at several typical locations.

RECEIVED R-F SIGNAL STRENGTHS AND PATH LOSSES

The first factor in evaluating mobile radio transmission is the strength of the r-f signal which is received. This is inversely related to the loss in the r-f path. The mobile units of a mobile system are either moving around or, if stationary, are located at random. Since the effects of the many geographical features, buildings, and the like, which influence propagation can combine differently for different locations of a car, even where the locations are only a fraction of a wavelength apart, the only meaningful measure of signal strength is a statistical one. Such statistical answers were obtained by making and recording many instantaneous samples of field strength with the aid of the LDR, mentioned above.

It is of interest to note that whenever the sample measurements were confined to a relatively small area, say 500 to 1000 feet or less in extent, the amplitude distribution of these samples tended strongly to follow along the particular curve known as a Rayleigh distribution. Such a curve and a typical set of experimental points are shown in Fig. 5. The same distribution was obtained at all of the frequencies tested, including 3700 mc. The rapidity of signal fluctuation, as the car moved, was proportional to frequency, but this does not affect amplitude distribution. Such a distribution could have been predicted if it had been postulated that the transmitted signal reached the car antenna by many paths having a random loss and phase relationship. It is thus inferred that in general the signal reaches a car by many simultaneous paths.

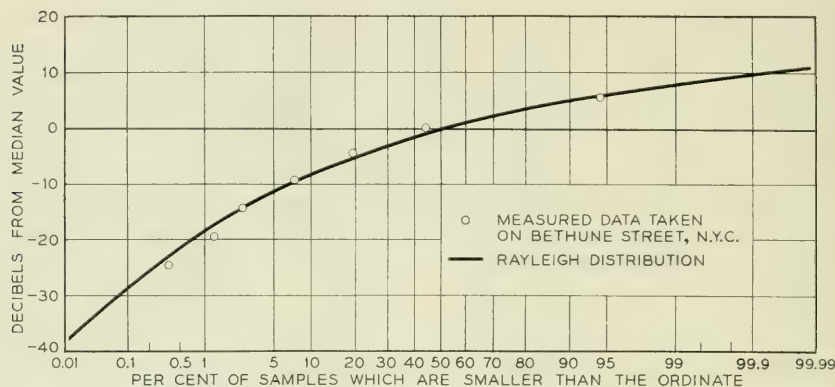


Fig. 5—Typical distribution of test samples of r-f signal strength taken over a small area.

With the shape of the distribution known, only one other value need be given in order to specify the propagation to such a small area. This might be the median, the average, the rms, or any single point on the curve. The one used most often here is the median, that is, the value which is larger than 50 per cent of the samples and smaller than the other 50 per cent. Measurement of the median value by this statistical method was found to be accurately reproducible, and therefore is presumed to be reliable. Successive batches of 200 samples each, all covering the same test area, yielded median values which differed not more than 0.5 db when none of the conditions changed; i.e., transmitter power, antenna gain, and receiver calibration remained the same. This accuracy may seem surprising when it is realized that individual samples differ frequently by 10 db, and often as much as 30 to 40 db.

It was presumed at the outset of the tests that the different frequencies would exhibit different propagation trends with distance. For this reason the samples have been grouped by distance. In presenting these results, it was convenient to express the measurements of received RF signal in terms of path losses. By this it is meant the loss between the input to a dipole antenna at the transmitter and the output of a whip antenna on the test car. These path losses will have, of course, the same distribution as the received r-f signal.

The results of the path loss measurements are given in Figs. 6, 7, and 8 for 150, 450, and 900 mc respectively. These values represent the loss between the input to a half-wave dipole antenna at one end of the path and the output of a quarter-wave whip at the other end. They are shown here as a function of distance from the land station. For distances under

ten miles the data are the result of tests in Manhattan and the Bronx. For each distance a test course was laid out approximately following a circle with that distance as a radius. The data for ten miles and greater distances were obtained on two series of tests along radials from the land transmitter, one of which followed Route 1 through New Rochelle, N. Y., and the other followed Route 10 toward Dover, N. J. For reference, a curve has been given on each of these figures which shows the computed loss based upon the assumption of smooth earth.

A curve labeled "1 per cent" means that in one per cent of the sample measurements the loss was less than that indicated on the ordinate. The meaning of the labels on the other curves is similar. The curve labeled "50 per cent" is, of course, the median.

It will be apparent that the assumption of smooth earth is not applicable to the area tested. The data for median losses are in the order of 30 db greater than the value computed over smooth earth. This additional loss may be thought of as a "shadow" loss arising from the presence of many buildings and structures.

The distribution of losses given in these three curves is wider than the Rayleigh distribution of Fig. 5. This is because the data for each

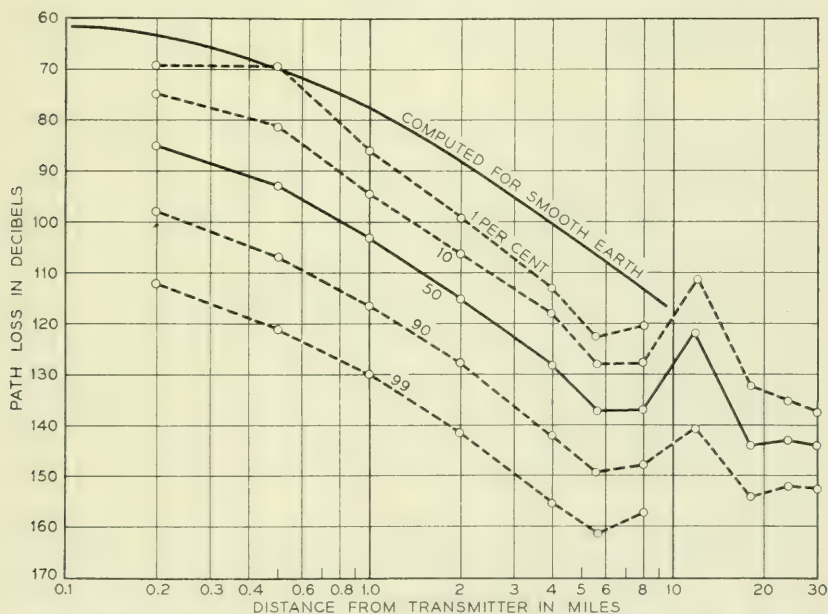


Fig. 6—Measured path loss at 150 mc in Manhattan and the Bronx and suburbs. (Note: Data for 10 miles and greater were taken on Route 1 toward New Rochelle and on Route 10 toward Dover.)

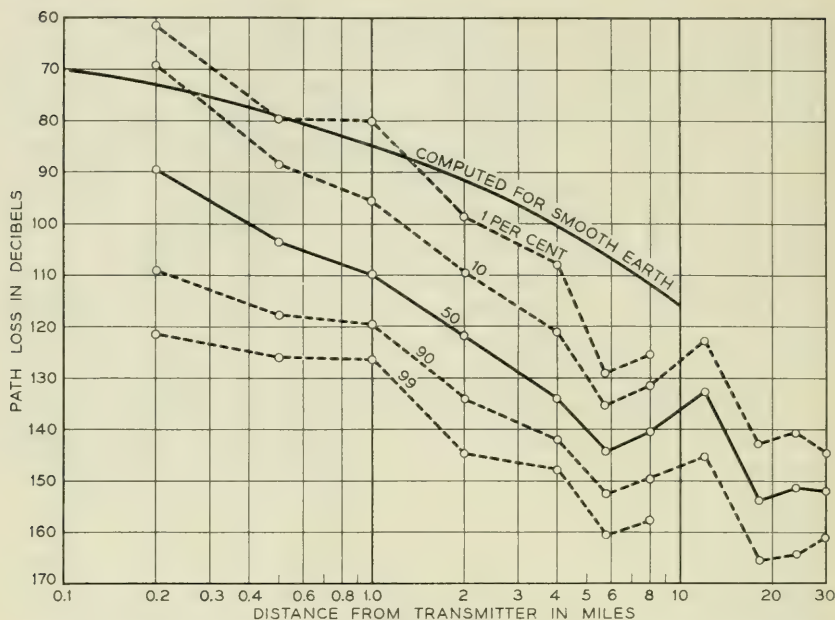


Fig. 7—Measured path loss at 450 mc in Manhattan and the Bronx and suburbs. (Note: Data for 10 miles and greater were taken on Route 1 toward New Rochelle and on Route 10 toward Dover.)

distance are a summation over many different locations rather than a set of samples covering one location.

The data for ten miles and further from the transmitter were taken on routes through suburban areas. The losses at twelve miles appear to be less than the average trend indicated by the curves. This is because data taken at the top of the First Orange Mountain weigh heavily at this distance. It is of interest to note that the losses at distances of ten miles and over are 6 to 10 db less than might have been predicted from the trend at smaller distances, where the measurements were made in city areas. This probably reflects the fact that there is a considerable difference in the character of the surroundings, such as height and number of buildings in the suburban territory as compared with the city itself.

The median curves of loss have been replotted for three frequencies on Fig. 9. This permits a better comparison with frequency. Except very close to the transmitter, the performance at the various frequencies seems to differ by an essentially constant number of db, while exhibiting the same trend with distance. The similarity between frequencies is appar-

ently much greater than the similarity between the median value and the value computed over smooth earth for any given frequency.

It was not possible to get complete enough data to plot a curve for 3700 mc similar to the ones mentioned above. The test setup at this frequency was limited by transmitter power and receiver sensitivity. Only those locations for which path loss was relatively low could be tested. A comparison of results at these locations is given in Figure 10. The curves labeled "1 mi.," "2 mi.," and "4 mi." for Manhattan are the median values obtained along test routes which followed circles of 1, 2 and 4 miles radius from the transmitters. The other curves refer to selected small areas at greater distances on the Hutchison River Parkway and New Jersey Route 10, as indicated. Although the data at 3700 mc not extensive, the trend with frequency seems clear.

More specific data for path losses measured along the routes toward Dover and New Rochelle are given in Fig. 11. Each value plotted here is the median of about 200 samples taken in a small area at the distance indicated. The strong effect of the First and Second Orange mountains at fourteen and sixteen miles on the Dover route is of interest.

The coverage desired in these mobile telephone systems extends into suburban locations. It follows that a comparison of coverage by the

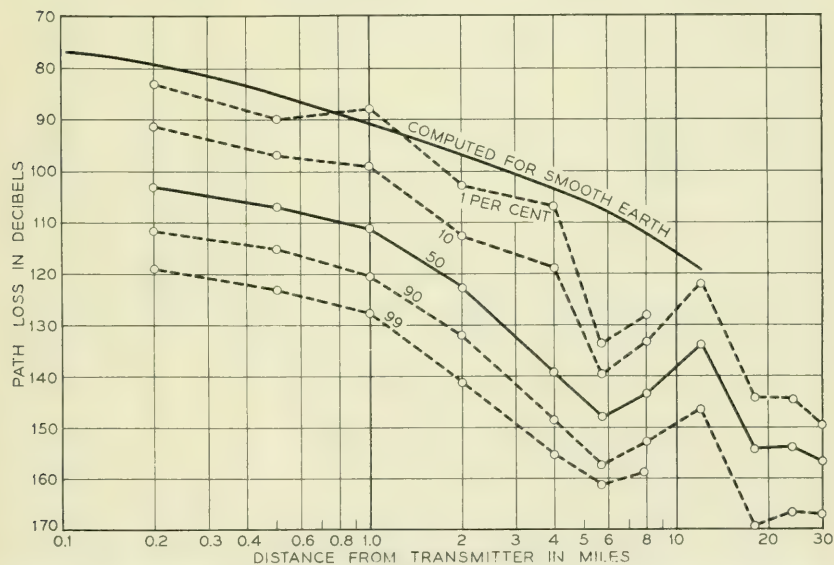


Fig. 8—Measured path loss at 900 mc in Manhattan and the Bronx and suburbs. (Note: Data for 10 miles and greater were taken on Route 1 toward New Rochelle and on Route 10 toward Dover.)

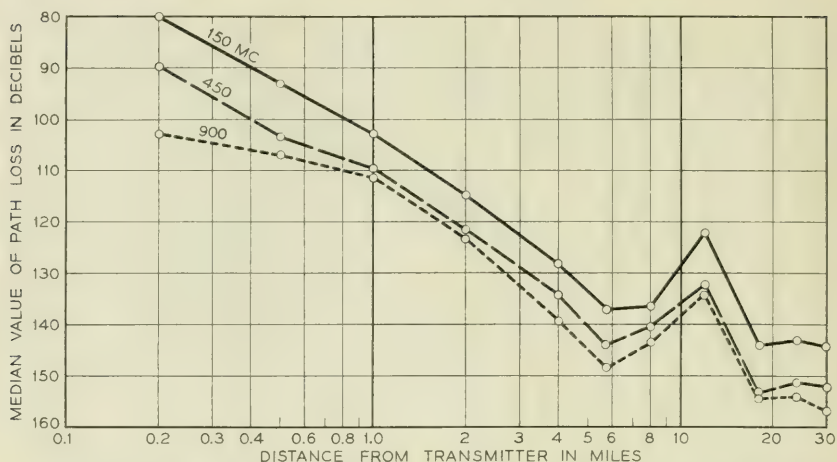


Fig. 9—Median values of measured path losses. (Note: Data for 10 miles and greater were taken on Route 1 toward New Rochelle and on Route 10 toward Dover.)

various frequencies should be based upon measurements taken in the suburbs. The data from the New Rochelle and Dover series have been used as a basis for the points and the curve given in Figure 1. Each of the circle points shows the path loss at a given frequency relative to that at 150 mc for a particular location. Their spread indicates that the

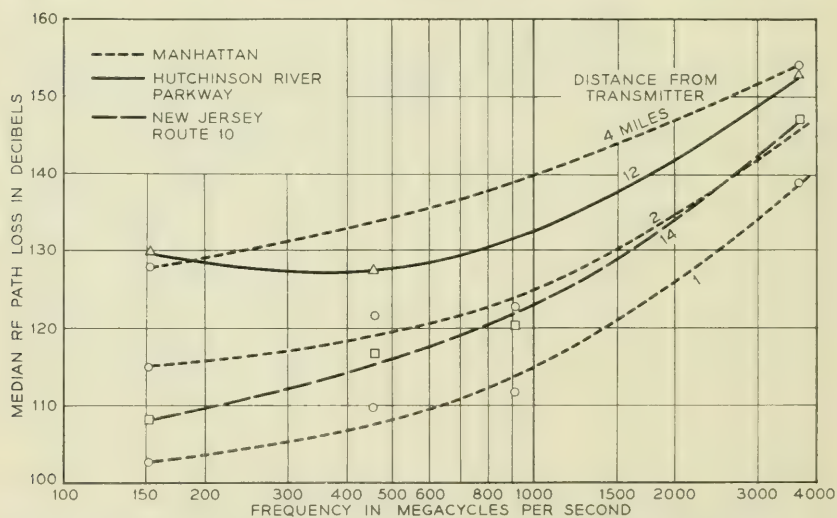


Fig. 10—RF path losses at locations for which 3700 mc measurements were made.

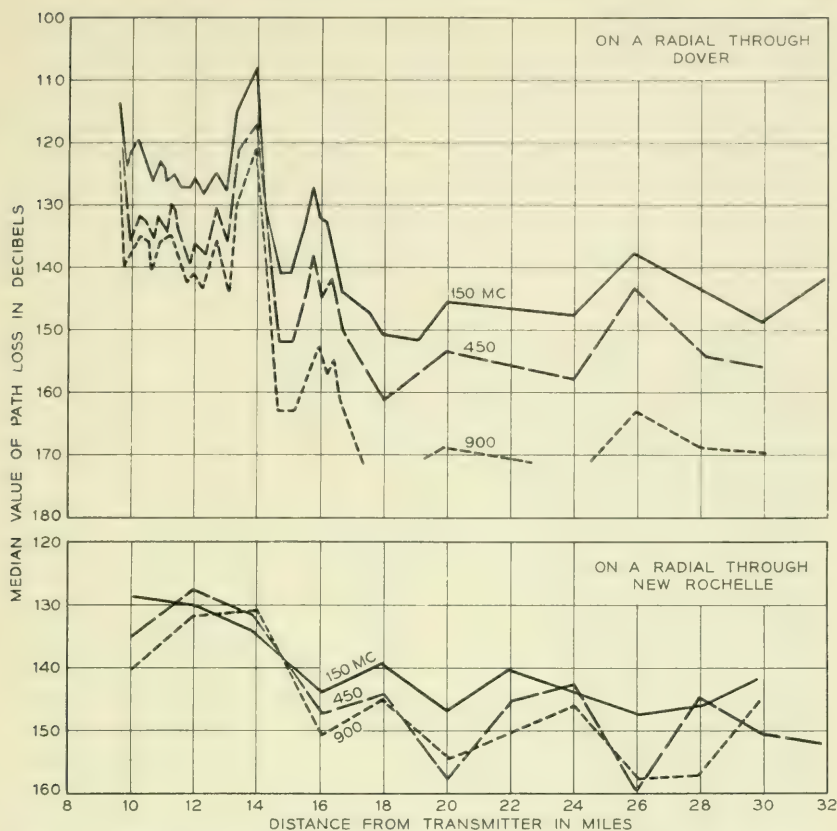


Fig. 11—Median r-f path losses along selected routes. (a) On a radial through Dover. (b) On a radial through New Rochelle.

comparison of frequencies is different at different locations. The “crosses” are the median values of these points, so placed that there are as many points above as below. The points for 3700 mc are taken from the data of Fig. 10. The crosses of Fig. 1 are considered to be the most reliable all-around comparison of propagation at the different frequencies.

RELATION OF SPEECH-NOISE RATIOS TO R-F SIGNAL POWER

Speech-to-noise ratios were measured at all of the test locations by the use of the level distribution recorder as described earlier. During the course of any given test the audio noise from the receiver varied considerably and these variations were recorded on the LDR. It was found by correlation between subjective observations of circuit merit and the

median value of noise that the latter is equivalent in noise effect to a steady random noise of the same value. In the FM receiver, the level of speech is essentially not affected by the strength of RF signal and so a measurement of the output noise is directly related to the speech-to-noise ratio. The speech-to-noise ratios given here are computed from noise measurements by assuming that speech of -14 vu level is applied to the system at a point where one milliwatt of 1,000 cycles tone would produce a 10-kc frequency deviation. The strength of the speech signal at the receiver output is expressed in the same units as are used for the noise.

As might have been expected the median speech-to-noise ratios correlate strongly with the amounts of r-f signal received at the various locations. This correlation has been evaluated in order that the most likely relationship between speech-to-noise ratio and received r-f signal may be known for the different frequencies. These are shown in Fig. 12, where each circle represents the median speech-to-noise value measured at one test location plotted against the median r-f signal received at that location. The solid lines have been drawn in to show the trend. The bending at the top of the curve is inconsequential. It only represents the limit imposed in the test setup by tube microphonic noise, vibrator noise, etc. The curves show, for example, that in order to produce a commercial grade of transmission, which requires a 12 db speech-to-noise ratio, the median r-f signal must be 122.5 db below one watt at 150 mc.

The data given in Fig. 12 pertain only to the suburban locations. Measurements in Manhattan have not been included, even though they indicate that larger signals are required, because the limit of system coverage is to be found in the suburbs. The data on the solid curves of Fig. 12 have been used to derive the curve of Fig. 2 which plots the value of r-f signal required at the mobile receiver for a commercial grade of transmission. The dotted curve of Fig. 2, which shows the median signal required in locations where noise picked up by the antenna is less than set noise, is based on the assumption of an 8 db noise figure for a practical 150-mc receiver, 11 db at 450 mc, and 12 db at 900 mc and higher.

Measurements have been made of the effect of noise picked up by the antenna at land receiver stations. These are expressed here in terms of the carrier strength required for just-commercial grade of transmission (12 db speech-to-noise ratio) as compared with the value required when there is no antenna noise and only receiver noise is present. These comparison measurements were made by injecting a steady carrier into the receiver with an antenna connected normally, and again with a dummy antenna connected. Although these tests were made with a steady rather

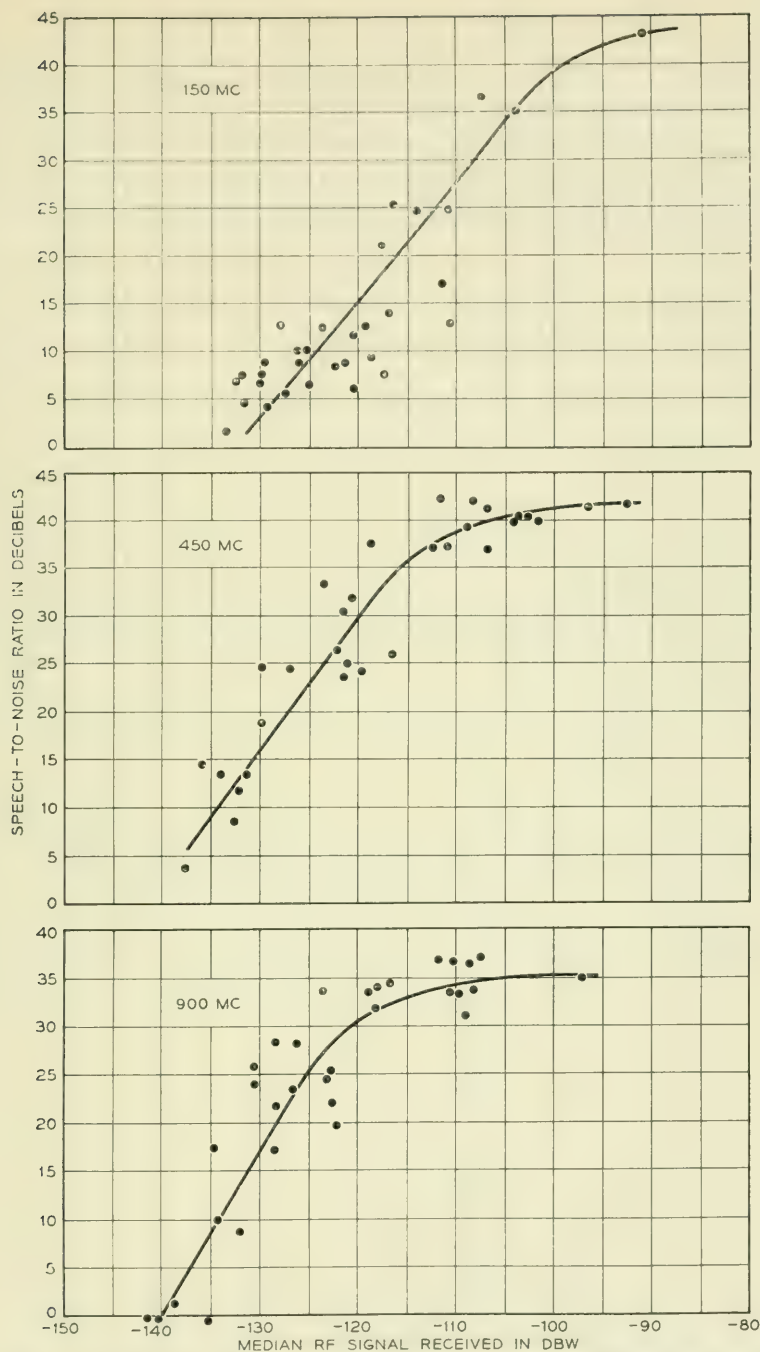


Fig. 12—Correlation between median values of speech-to-noise ratio and r-f signal strength in suburban locations. (Note: Each point represents the speech-to-noise ratio and the r-f signal received at one location. (a) 150 mc. (b) 450 mc. (c) 900 mc.

than randomly varying signal, it is felt that the comparative results will apply to the random signal case as well.

Tests were made at 150, 450 and 900 mc, at four locations of interest, and with dipole and 7 db gain antennas. Not all combinations were tested, but enough to permit some interesting comparisons. The locations tested were as follows:

A: On the Long Lines building, a 27-story building in downtown Manhattan.

B: On the Graybar-Varick building, a 16-story building in downtown Manhattan.

C: On the telephone building which houses the Melrose exchange, a 7-story building in the center of the Bronx.

D: On the 3-story telephone exchange building in Lynbrook, Long Island.

Table I describes the generally prevailing noise situation at these locations. Higher noise was encountered occasionally at some of the sites, due in at least one case to operation of elevators in the building. However, these occasions were so brief and infrequent that the general background of noise is considered to be a better value to use in estimating systems performance.

The trend toward lower site noise at higher frequencies, already noted for mobile installations, is seen to apply to land receivers as well.

TABLE I—R-F SIGNAL INPUT TO RECEIVER FOR 12 DB SPEECH-TO-NOISE RATIO (GIVEN IN DB ABOVE THAT NEEDED TO OVERRIDE SET NOISE*)

Station	Frequency	Antenna	
		Dipole	7-db Gain
A	150 mc	10	—
	450	1	0.5
	900	—	1.5
B	150	12	—
	450	4	—
	900	0	—
C	150	11	2.5
	450	1	1
	900	1	—
D	150	5	4
	450	0	0
	900	0	—

* Noise figures in the test receivers were 9, 12 and 12 db for 150, 450, and 900 mc, respectively.

These data bring out another interesting and significant fact. Where noise collected by a dipole antenna is discernible over set noise, the noise collected by the 7 db gain antenna at the same site is, surprisingly, less. This means that the gain antenna picks up *less* noise power than a dipole. Since it picks up 7 db more signal from a distant car, a gain antenna thus provides a double improvement in transmission at those sites for which ambient noise is controlling.

An explanation of this behavior may be surmised if it is assumed that the sources of noise are numerous and are scattered around at street level (motor vehicles, mostly). The overall noise received is a sum of contributions from all sources, weighted for distance and the receiving antenna pattern. A gain antenna of the type considered here tends to ignore the strong nearby noise sources because they are below the antenna beam. The sources, which are nearly enough in the beam to count, are also further away and are attenuated by distance.

The amount of data given in Table I does not seem sufficient to warrant stating a firm figure as to the amount of improvement obtainable from a gain antenna. However, substantial improvement at 150 mc is indicated, and this might have the effect of bringing the value of mobile transmitter power required at 150 mc down to the value required at 450 mc, assuming gain antennas in both cases.

ACKNOWLEDGMENTS

A number of people participated at one time or another in setting up and carrying through these tests. It is not possible to name them all, but the principal participants were R. L. Robbins, R. C. Shaw, W. Strack, D. K. White, and F. J. Henneberg. The program was supervised by D. Mitchell. The special radio equipment required was designed and furnished by W. E. Reichle and his group.

Common Control Telephone Switching Systems

By OSCAR MYERS

(Manuscript received August 1, 1952)

In the development of dial telephone switching systems two fundamentally different arrangements have been devised for controlling the operations of the switches. In one arrangement the switch at each successive stage is directly responsive to the digit that is being dialed. Systems using this method of operation are called direct dial control systems, an example being the step-by-step system as commonly used in the Bell System. In the other arrangement the dialed information is stored for a short time by centralized control equipment before being used in controlling the switching operations. Systems using the second arrangement are known as common control systems, examples of which are rotary, panel and crossbar. These two arrangements have different economic fields of use, the direct dial control being better suited for the smaller telephone exchanges and the common controls for the larger exchanges, especially those in metropolitan areas. A history of the evolution of these types of switching systems is presented, followed by a discussion of their comparative merits for various fields of use.

HISTORY

Invention of machines for switching telephone connections started shortly after the invention of the telephone. A forerunner of the step-by-step system, the Connolly and McTighe "girless" telephone system,* was patented in 1879 and the first patent on the Strowger step-by-step system† was issued in 1891. The first commercial installation of automatic switching equipment was made at La Porte, Indiana, in 1892. This installation used step-by-step mechanisms.

In the early 1900's many telephone engineers regarded full automatic switching as uneconomical but technically feasible if restricted to single office exchanges with individual flat rate lines. They were, however, un-

* U. S. Patent 222,458—1879—Connolly and McTighe.

† U. S. Patent 447,918—1891—Almon B. Strowger.

certain about the future of this method of operation. It appeared to them that the greatest promise in the use of automatic apparatus was in distributing calls to manual "A" operators and in the elimination of the "B" operators. Consideration was being given to systems capable of operating on either a semi-mechanical or a full mechanical basis depending on whether the dial was located at the "A" board or at the subscriber's station. Development was also under way to provide arrangements for trunking calls between dial offices and to overcome the numerous weaknesses and deficiencies of existing dial systems.

The Strowger Company, the Bell System, and several other companies were planning or developing automatic and semiautomatic systems at that time. These included the full automatic, the network automatic, the automatic operator, and the semiautomatic. Short descriptions of some of them follow.

EARLY FULL AUTOMATIC SYSTEMS

The full automatic systems were mostly direct dial control. They included the Strowger, the Western Electric 100-line and 20-line, the Clark, the Faller* and the Lorimer systems.

The Strowger system of the middle 1890's provided 100-point two-digit selectors, one for each line. For each group of 100 lines the 100 outlets of each selector were multiplied to the corresponding outlets of the other selectors serving the group. Each outlet of the group ran to a two-digit connector, each connector having access to 100 lines. Thus every group of 100 lines had 100 selectors and a maximum of 100 connectors and could reach 10,000 lines in a full office. Each group of connectors, up to the maximum of 100 connectors per group, had a multiple of 100 terminating lines. This was therefore a 4-digit single-office system theoretically of 10,000 lines capacity, requiring 1 selector and 1 connector per line. Subscribers in a given originating group of 100 lines had only one path to a particular terminating group of 100 lines. Since a selector was provided for each line, no dial tone was necessary. The switches used the familiar up and around motion. The exchanges of this type that were installed were small, the largest being in the order of 1000-line capacity. This type was followed by a new arrangement when automatic trunk selection was introduced. This provided multiple paths to each terminating group of 100 lines; the selector at this stage became a single-digit switch.

The Western Electric 100-line system could actually serve only 99

* U. S. Patent 686,892—Ernest A. Faller—Nov. 19, 1901.

lines. (The record does not disclose why one of the terminals of the system could not be assigned.) It used a rotary selector per line directly driven by a single train of pulses generated by a lever operated dial at the station. The selector had 100 points and the number of pulses sent corresponded to the number of the called line. The 20-line system was similar to the 100-line system.

The Clark system was a single motion rotary step-by-step system using 75-point switches which accommodated a maximum of 74 lines. (Here again there is no record as to why one terminal was not used for a line.) It did not provide a busy test. There were no relays in this system.

"AUTOMATIC OPERATOR" SYSTEMS

The Faller and the Lorimer systems were called "automatic operator" systems but they were actually versions of direct dial. The Faller system was apparently never used commercially, but the Lorimer system was.

The inventors of the Lorimer system had several objectives. One was to produce a system which could be installed in 100-line building blocks, called sections. As little as one section could be installed and operated alone. Additional sections in increments of 100-line capacity could be added as required up to the limit of 10,000 lines. Another object was to get good contacts and they therefore employed switches with heavy contacts like those used in power switches. The power needed to drive switches with such contacts led to the adoption of a common power drive for a number of switches instead of electromagnets individual to the switches. Still another aim was to provide a minimum of equipment on a per line basis and to provide equipment only to the extent required by traffic. Line relays were therefore omitted in early offices and the 100-line sections were divided into divisions, maximum 10 divisions per section, with arrangements for omitting divisions if not required by traffic.

The Lorimer system was a direct dial system operated from a pre-set calling device. It had a line finder stage, a selector stage and a connector stage. The calling device, wound up by a crank, had four settable levers, one for each digit, each of which grounded one terminal in its own set of ten terminals corresponding to the digit set up. The levers also operated a visual indicator. In the calling device there was also a switch driven over its terminals by a magnet-controlled escapement. Pulses were sent from the central office to control the escapement and the central office equipment was driven in synchronism with the station

switch until a grounded station terminal was found. The central office equipment was then stopped but the station switch continued stepping until the starting point for the next digit was reached. When the central office equipment was ready for the next digit the process was repeated until the called line was reached.

The Lorimer system has now disappeared from the scene in spite of a number of attractive features. The reasons for this disappearance are not clear from available records, but some reasonable conjectures can be made. For one thing, the pre-set calling device must have been expensive both in first cost and to maintain; it was also designed for a maximum of four digits and a re-design for more than four digits would have entailed substantial effort for developing both the calling device and the central office equipment. There is also some evidence to indicate that the system cost more than either step-by-step or panel.

THE NETWORK AUTOMATIC SYSTEM

The network automatic was a proposed form of semiautomatic in which the subscribers retained their manual instruments and were served by small unattended branch offices, each of which had a single group of trunks to a central operator office. On originating calls the branch offices acted as concentrators, automatically connecting calling lines to trunks to the central office where the operators were located and who asked for the called number as in straight manual practice. Called lines were reached through the branch offices by the operators at the central office who were provided with keysets to control the branch office equipment.

SEMI-AUTOMATIC SYSTEMS

There were several plans for other types of semi-automatic systems. Most of them contemplated replacing the "B" operator by a machine under control of the "A" operator. The plan of using machines under control of the "A" operators to replace the "B" operators was operated successfully in Saginaw, Mich. with Strowger apparatus. A similar plan was in operation in Los Angeles, and several groups of engineers studied improvements and variations.

STATUS IN 1905

The status of automatic switching by 1905 was this: there were several single office cities which had commercial installations of Strowger step-by-step equipment with severe limitations even for this field of use; a

number of Western Electric Company 100-line and 20-line automatics were in commercial service; a small amount of semi-automatic equipment was also in operation with the equipment under direct control of the "A" operator's dial; and planning and development work were under way to remove some of the limitations and extend the field of use of the automatic and semi-automatic systems.

The rotary dial was developed in 1896. However, many of the early systems did not use this type of dial. Various calling devices were used for a number of years. Among these were lever operated pre-set devices, keysets of several types, and dials with holes (in one case as many as 100) in which a peg could be inserted to act as a stop for an arm which was pulled around and allowed to restore. In all the early systems, regardless of the device used, the signals generated at the calling station directly controlled the selections.

RECOGNITION OF NEED FOR ACCESS TO LARGER TRUNK GROUPS

While mechanisms and circuits were being developed for direct dial control switching, work of a theoretical nature was going on which was to have an important effect on future designs. This work consisted of traffic probability studies and observations the outcome of which was the development of formulae and curves on the efficiency of trunk groups which influenced strongly the views of engineers as to the economical sizes of switches. G. T. Blood of the American Telephone and Telegraph Company in 1898 found that the binomial distribution closely fitted the observed data on the distribution of calls. The first comprehensive paper on the matter was one by M. C. Rorty in 1903, *Application of the Theory of Probability to Traffic Problems*. Curves accompanying his paper indicated that trunking efficiency improved with group size. Subsequent work by E. C. Molina in postulating that the grade of service experienced by a particular call applied to every call in the office and in developing the Poisson approximation to the binomial expansion formed the basis for trunking theory as used in the Bell System. Fig. 1 is a reproduction of three curves produced by Molina on July 6, 1908, showing the average load carried by various numbers of trunks for three probability conditions namely P.01, P.001 and P.0001 corresponding to an all trunks busy condition encountered by calls once in a hundred, once in a thousand, and one in ten thousand times respectively. From these curves it can be seen, for example, that ten trunks can carry a load averaging slightly over four calls with a probability of loss of P.01. Twenty trunks can carry an average of over eleven simultaneous calls

with the same P.01 loss but with an increase of efficiency of 15 per cent. The efficiency rises from 41 to 56 per cent.

EVOLUTION OF PRINCIPLE OF TRANSLATION

These studies had considerable effect on the trend of system design. For example, it appeared that grouping subscriber lines on the connectors in groups of more than 100 might result in some economy and that other economies were possible if the limitations imposed by decimal selections were avoided.

However, a new invention, namely translation, was required before systems could operate with large access switches and non-decimal selections. Translation is a mechanical rearrangement which permits conversion of the decimal information received from the dial to non-decimal forms for switch control and other purposes. When translation is made changeable by some means such as cross-connections, it is the basis of much of the flexibility of common-control systems. Translation was first proposed by E. C. Molina late in 1905. A patent application* for a *Translating and Selecting System* was filed on April 20, 1906.

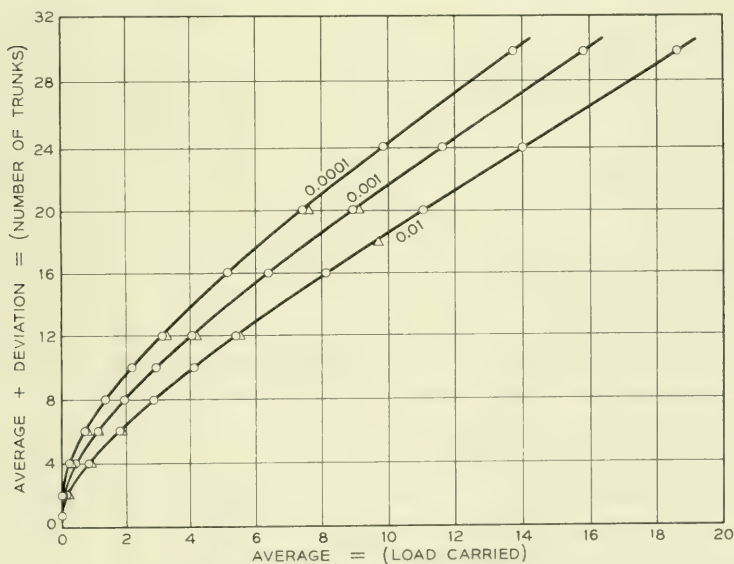


Fig. 2—Bypass system.

* Patent No. 1,083,456 issued to E. C. Molina, Jan. 6, 1916.

A necessary feature of systems employing translation of a series of digits such as an office code is digit storage. It was only a small step from the concepts of translation and digit storage to arrangements which provided these features in common circuits. Common controls with translation were first employed in the rotary system.

THE ROTARY AND PANEL SYSTEM DEVELOPMENTS

The rotary system was a full-fledged common-control system using register-senders to store the dialed information, to translate it to control the two-hundred point ten-level power-driven switches in selecting outgoing trunks from the originating office and in making line selections in the terminating office. The translation of the digits used for selecting trunks was changeable, but the translation of the numerical digits was fixed in permanent wiring of the register-senders.

In a search for less expensive cabling arrangements than those required by the rotary system, the panel bank employing punched metallic strips was developed. Each bank in the selectors of this system can accommodate 100 outlets with three wires per outlet, and five banks are stacked into a frame over which 60 power-driven selectors can hunt. For several years, starting in 1907, parallel development of the rotary and panel systems was carried on and desirable features of one were incorporated in the other. The panel system also has register-senders with changeable translation for selecting trunks and fixed translation for controlling selections in the terminating equipment. The major differences in the early designs of rotary and panel were due to the different access of the two systems and to differences in the methods of controlling the selectors. Both panel and rotary use reverive pulsing to control the selections. With reverive pulsing as the selectors progress they send back pulses which the sender counts. When a selector reaches the desired position, the sender stops it by opening the pulsing circuit. Both panel and rotary, like the Lorimer system, use a continuously operated power drive common to a number of switches because the increased size of switch which the greater access of these systems required, made a separate power drive economical.

The panel and rotary systems were originally designed for semi-mechanical operation with automatic distribution of calls to operators as a possible adjunct and with provision for full automatic operation if it proved desirable, by locating the dial or some other calling device at the subscriber's station rather than at the operator's position. This was a reasonable plan when development of these systems was started. Studies indicated that semi-mechanical systems could reduce the number of

operators required by an amount ranging from 30 to 50 per cent by eliminating the "B" operators and increasing the efficiency of the "A" operators. At that time, full automatic systems were subject to a number of shortcomings such as the complications and unreliability of the pulsing device at the subscriber's station, the need for a local battery at the station, and the lack of arrangements for party line and message rate service. Furthermore, there was considerable doubt as to the ability of the subscriber to dial with acceptable accuracy the six or seven numerical digits required in some of the multi-office exchanges.

There was an acute need for relief from the difficulties of manual operation after the start of World War I. Telephone growth was so rapid that it appeared for a time that the demand for new operators, particularly in the large cities, might outstrip the available supply. Competition from other industry for female help was also increasing. As more offices were added, the situation was further aggravated by the increasing complexity of operation. On account of the increasing number of trunked calls, the growing number of central offices, and the increasing amount of manual tandem operation, the quality of service was being degraded.

DEVELOPMENT OF A LARGE CITY NUMBERING PLAN

By 1916, the full automatic system (Strowger) had established a competitive position with manual for single-office cities, and both manual and full automatic offices were considered to be more economical than semi-mechanical for such cities. Because the number of dial pulls for a single office was four or less, little concern was felt about dialing accuracy.

For the multi-office cities it appeared that full mechanical operation would improve service and be more economical than either the semi-mechanical system or manual and would reduce the pressing need for operators. However, in spite of these factors urging the adoption of a dial system and even though automatic equipment was actually used in Los Angeles and Chicago in the first decade of the century, there was a reluctance to adopt full automatic operation in the very large multi-office cities because of the lack of a suitable numbering plan. A cumbersome plan was under consideration for handling dial traffic in these cities. This required the use of seven-digit numbers with the dial customers being called on to use arbitrary three-digit numerical codes for the office names. At the same time, the existing office names would be retained for use by the manual customers. Adoption of this dual arrangement would have required the provision of a cumbersome directory, but worse than that, it was felt that dialing seven numerical digits would be too

confusing to customers and that consequently there would be an excessive number of dialing errors. It was therefore planned to use semi-mechanical operation for cities like New York, retaining an operator between the customer and the machine. While this scheme did not save as many operators as the full mechanical method, it was believed necessary to have trained operators so that the customers would not be subjected to the complications of dialing. Under the proposed arrangement, the customer would pass the office name and number orally, and the operator would substitute the dial code for the office name and key or dial the code and number into the machine. Trial installations of the semi-mechanical panel system placed in service in the Waverly and Mulberry offices, Newark, N. J., in 1915 demonstrated that this method could provide reliable and improved telephone service under severe conditions.

However, in 1917 W. G. Blauvelt of the American Telephone and Telegraph Company proposed a numbering plan which would permit the customer to dial up to seven digits with acceptable accuracy and which would also be satisfactory for manual operation. This arrangement consisted of the use of one to three letters and four numbers. The first one, two or three letters of the office name were printed in bold type in the directory as an indication to dial customers that these were to be dialed ahead of the four numbers. Manual customers used the office name as before. Letters as well as numbers were placed on the dial plate in line with the finger holes of the dial. This proposal was immediately adopted and further Bell System development proceeded along the lines of full automatic operation. The Bell System planned to use the panel system in large cities not only because of the trunk efficiency which was possible with the use of the large panel switch, but also because trunking, being no longer under direct control of the dial in this system, was divorced from numbering. The panel system was also attractive because it had flexibility for growth and for contingencies such as the introduction of new types of service. These advantages would be provided by the common senders and translators of that system.

EARLY INSTALLATIONS OF COMMON CONTROL SYSTEMS

Early in 1918 tentative schedules were set up for 6-digit panel offices for Kansas City and Omaha and late that year a 7-digit office was recommended for the Pennsylvania office in New York City. When the Atlantic office in Omaha was placed in service on Dec. 10, 1921, it became the first commercial installation of a full automatic panel system.

Commercial installations of rotary equipment preceded the first com-

mercial panel offices. A semi-mechanical rotary system was installed in Landskrona, Sweden, in 1915 but remained in service for only a short time. A similar system was installed later in 1915 in Angiers, France. The first full mechanical rotary installation was at Darlington, England, in 1914. This system is still in service.

A common control system using Strowger switches, the director system, was developed in 1922. This development was prompted by the desire to provide automatic equipment in the London, England, multi-office exchange where the layout of the outside plant required considerable tandem trunking if a reasonably economical trunk network was to be achieved. All of the outside plant in London for the manual system was underground and it was required that this arrangement be retained when dial equipment was installed. This tended to fix the routes of telephone cables and to make it expensive to open new direct routes as new offices were opened. The trunking economies of tandems were extremely desirable under this condition and common controls with translation were necessary for a practical scheme capable of operating with the tandems. The director scheme, which in principle parallels the sender-translator scheme of the panel system, was designed to meet this situation. The director system was first placed in operation in Havana, Cuba, in 1924 and later in London in 1927.

EVOLUTION OF THE MARKER PRINCIPLE

In retrospect, it is obvious that the development thinking up to the early 1920's was limited by the belief that it was necessary to have the selectors do the testing for idle trunks even with common controls. This arrangement had been successfully used in the step-by-step system and it was natural to follow the same plan in the panel, rotary and director systems. Subsequent development of the common-control idea, starting with an experimental "coordinate" system in 1924, has resulted in marker systems in which the trunk testing is done by the markers.

The coordinate system derived its name from the method of operation of its switch, the process resembling the method of marking a point by the use of coordinates. The switch was essentially a large version of the crossbar switch and selected and held a set of crosspoints by the operation of horizontal and vertical members. Translation of the called office code, selection of a trunk, and operation of the switches to connect a transmission circuit to the trunk were functions of a new circuit, the marker, which the sender called into use for a fraction of a second after it had received the office code digits.

When the marker does the testing for idle trunks the trunk access from

a particular switch is no longer a limiting factor in the size of the trunk group. Once markers were invented it became possible to design systems using markers to do the trunk testing and any type of switch to do the connecting. When a trunk has been selected by the marker, the appropriate switches can be operated to connect to the marked terminal. The maximum size of trunk group need not be limited by the number of terminals on one switch. With a primary-secondary switch array groups much larger than those accessible on a single switch can be handled.

The coordinate system was not developed for commercial use. The first commercial marker system was PResident 2, a No. 1 crossbar office cut into service in Brooklyn, New York, in February, 1938. Improved crossbar systems have been developed since then including No. 5 crossbar and several types of toll crossbar systems.

There is an interesting sidelight on the development of crossbar systems. The crossbar switch was invented by J. N. Reynolds of the Western Electric Company in 1913.* At that time proposed plans for using this switch assumed that it would be used as a line switch. The arrangements did not appear attractive and no serious attempt was made to develop a commercial system using the switch either as a line switch or as a selector. A number of years later an improved version of the crossbar switch was developed by the Swedish telephone administration. Their plans contemplated the use of the switch as a selector in a direct dial control system. In 1930 W. H. Matthies of Bell Telephone Laboratories visited Sweden and, impressed with the possibilities of the switch, ordered samples from Sweden after his return to the United States. Work was started to improve the switch and to develop a modern system around it. The crossbar switch, as previously mentioned, was a small version of the coordinate switch and the development of No. 1 crossbar was therefore started on a plan which was based on principles used in the coordinate system some of which had been successfully applied to the panel system with the adoption of the decoder in 1927.

TYPES OF COMMON CONTROL SYSTEMS

Four basic variations have been used in systems with common controls. These are (1) digit storage in common circuits on a decimal basis and control of switches by the stored digits without translation; (2) digit storage in the common circuits on a decimal basis, fixed translation and control of switches in a fixed pattern by the translated information; (3) a modification of the preceding plan in which the translation can readily

* U. S. Patent No. 1,131,734—J. N. Reynolds—issued March 16, 1915 and re-issued December 26, 1916.

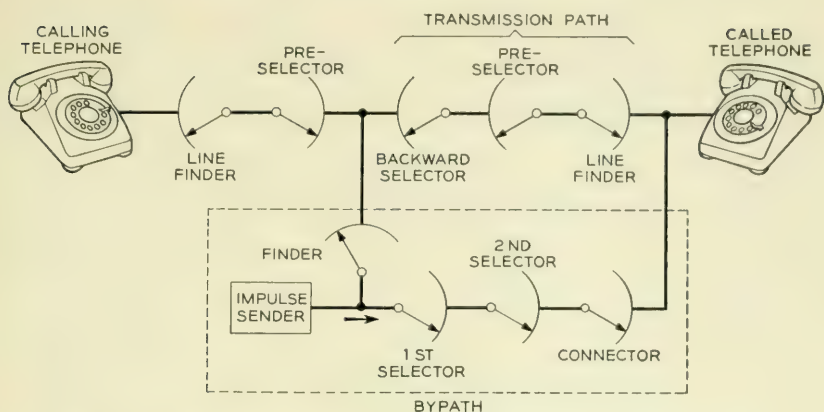


Fig. 1—Curves developed by E. C. Molina for trunk engineering.

be changed for any item of traffic; and (4) a still further variation where the function of hunting for an idle path is removed from the selectors and placed in new circuits called markers. Each variation resulted in improvements over preceding methods of operation.

The first plan is the simplest but also the least flexible. An advantage of this arrangement as well as of the other plans which also store the digits over step-by-step is that the interdigital time does not control the group size. By-path systems are examples of this method of operation. A system of this type is shown in Fig. 2. By-path systems use an auxiliary switch train that is under direct control of the dialed pulses to set up a connection. The talking circuit is then established over a parallel system of switches. The auxiliary train releases after the talking connection is set up and is available for use in setting up other connections. The Lorrimer system avoided the penalties resulting from hunting during the interdigital interval by storing the digits at the station.

A further step in the direction of flexibility, but with added complication, can be taken by a *fixed* translation from a decimal to a non-decimal basis, i.e., a form of translation wherein a given decimal digit or a set of decimal digits is always changed into the same predetermined non-decimal equivalent. This permits the use of switches with less than ten groups of outlets thereby providing economies by permitting larger groups of outlets with a given size of switch.

A third variation with still greater flexibility than the first two, but also with greater complication, is a system with *changeable* translation. Changeable translation is achieved by providing some means such as cross-connections for readily changing the output pattern of the translators generally for sets of digits as, for example, for the called office

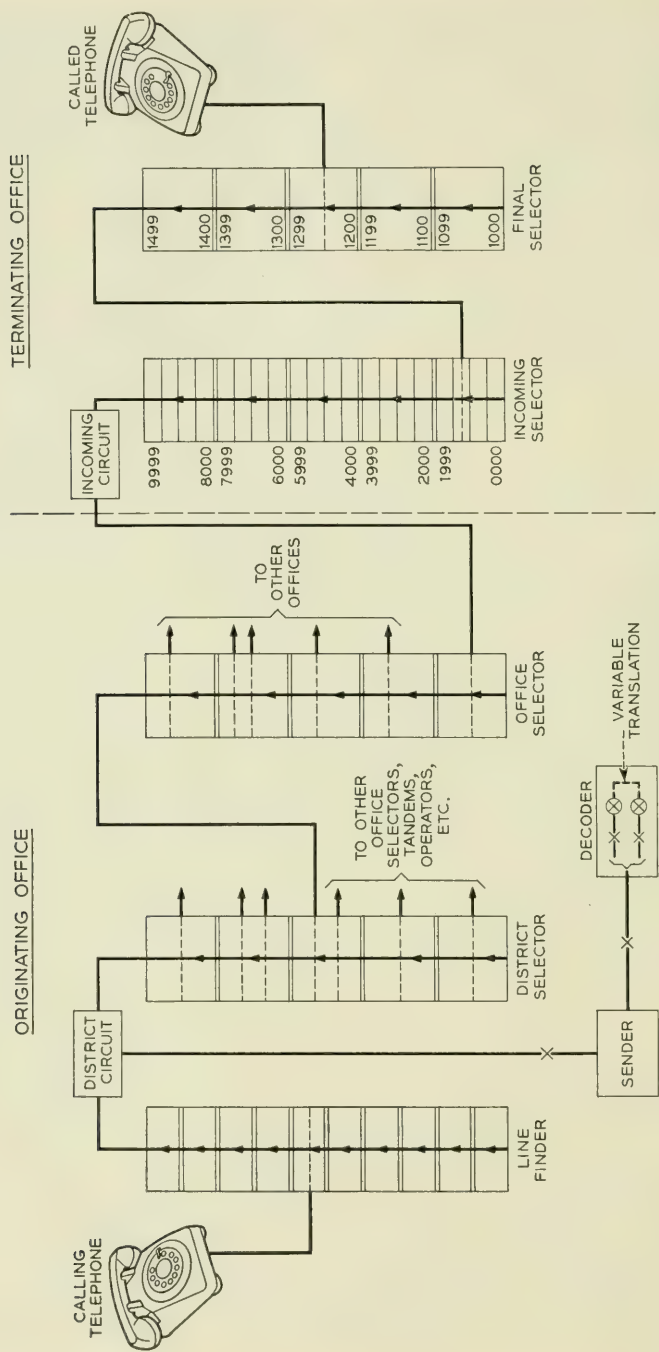


Fig. 3—The panel system.

codes. Changeable translation of office codes removes the limitation that the trunks for a given office designation must be located in a definite position on the switches which is the necessary result of fixed translation. Increased flexibility of numbering is now possible because office designation changes no longer require rearrangements of switch multiple. More economical arrays of switches are also possible because the switching plan can conform to traffic requirements without regard to numbering. Other advantages of translation—and as a practical matter, flexible translation—include the ability to operate with tandems, to operate with more than one type of outpulsing, and to operate with varying numbers of digits. The originating equipment of the panel system is an example of a system using changeable translation. This type of translation is also used for called line numbers as well as office codes in No. 1 and No. 5 crossbar thereby permitting these systems to shift lines for load balancing purposes without requiring numbering changes.

Finally, there is the most flexible but also the most complicated plan of all in which the selection of paths and trunks or lines is divorced from the selectors and placed in markers. In this plan the size of group is not limited by the number of terminals that a switch can hunt over in one sweep. No. 1 crossbar is an example of a system using the marker method of operation. In this system a switch generally has access to only ten trunks but on any one call a marker can test 160 trunks distributed over a number of switches.

Typical common control arrangements for systems using translation are shown in Fig. 3 for the panel system and in Fig. 4 for No. 1 crossbar.

The advantages noted are, in each case, the fundamental ones. Many others are inherent in common control and some will be brought out in further discussion.

A number of common control systems embodying the principles discussed have been designed. Rotary, panel and coordinate have been previously mentioned. Although the coordinate system never reached the commercial stage as a complete system, some of its features were adopted in the panel system starting in 1927 with the introduction of the decoder to replace the original three digit panel translator which used special panel selectors and pulse generating drums to do the translating job. This translator was limited in the digit combinations and number of three digit codes it could handle and also demanded a great deal of attention by the maintenance force. In place of the panel translators a small group of all-relay decoders, ranging from three to six, depending on traffic, was provided for each office. Senders were connected to decoders for about one-third of a second per call to obtain the information derived

from translation of the three office code digits. The connector for making the momentary connection of the large number of leads required between the senders and decoders presented new problems which were solved by the development of new relay preference and lockout circuits to permit as many simultaneous connections between senders and decoders as there were decoders and to permit an even distribution of calls to decoders. Decoder circuits were completely self-checking for trouble, provided for second trial in another decoder when trouble was discovered, and recorded troubles on a lamp bank trouble indicator.

In the early 1930's, encouraged by the success of decoders, the Bell System started development of the No. 1 crossbar system with markers in both originating and terminating equipments and with improved features over the coordinate system which it resembled in many respects. Self-checking circuits, second trials and trouble indicators which had proven highly successful in the decoder type panel system were important features of No. 1 crossbar. Automatic alternate routing and the ability to operate with non-consecutive PBX assignments were major new features introduced in this system for the first time.

The subsequently developed No. 5 crossbar system included a number of improvements, the chief of which from a common-control standpoint was the use of common markers for originating and terminating business and the use of the call back feature in setting up the connection. In this system the common equipment records the calling line identification as well as the called number, and after setting up to the called line or outgoing trunk, breaks down the connection to the common equipment

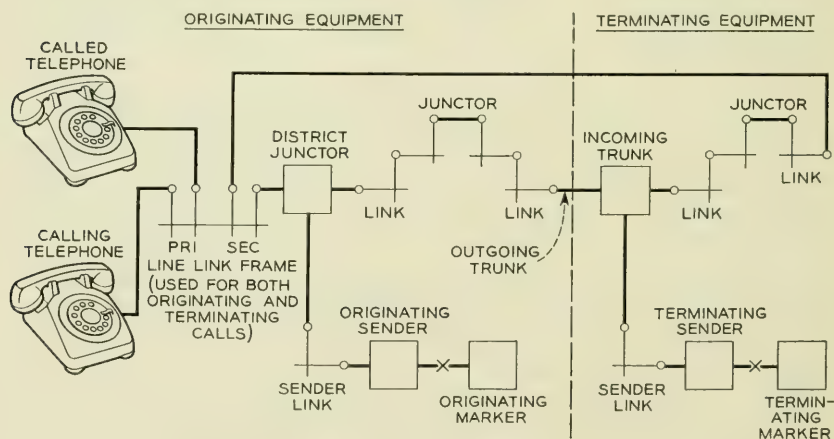


Fig. 4—No. 1 crossbar.

from the calling line and then re-establishes a connection back to the calling line.

Common controls have been employed by the Bell System in a number of systems in addition to those already mentioned. These include panel sender tandem, crossbar tandem, and No. 4, A4A and 4A toll crossbar.

COMPARISON OF COMMON CONTROL SYSTEMS AND DIRECT DIAL CONTROL SYSTEMS

Both direct dial control and common control systems have been developed to meet a wide range of situations for both large and small exchanges but, as previously noted, direct dial control systems have found their greatest field of use in the smaller exchanges and common control systems in the larger ones. The reasons for this can be brought out by a discussion of some of the features which have an important bearing on costs. These include the features affecting numbering plans, trunking arrangements, flexibility, quality of service, maintenance and engineering. A discussion of all the factors affecting costs will not be attempted. However, some of the more important ones will be covered.

RELATION BETWEEN TYPE OF SYSTEM AND NUMBERING PLANS

The requirements of a good numbering plan are well known. A good plan must be universal, i.e., must use the same number for reaching a called line regardless of the point of origin of the call in the area covered by the numbering plan, must permit dialing with acceptable accuracy, must permit directory listings that are readily understood by both dial and manual customers, and should use a minimum number of digits to reduce the labor of dialing. In small networks a satisfactory plan can be set up with almost any kind of system. However, especially in large networks, modern common control systems have outstanding advantages with respect to numbering.

These advantages of common controls are derived from the more flexible method of operation. Direct dial control systems use up the digits in the various stages of the switching operations whereas common control systems momentarily store them and can retransmit them. The result is that where direct dial control systems are used the numbering plan and the switching and trunking plans must conform whereas with common controls numbering, switching and trunking are not directly dependent on each other because the digits can be stored and translated. The effects of these differences on permissible latitude in numbering arrangements can be brought out by some examples.

Direct dial control systems cannot operate economically with a universal numbering plan in a network requiring any given call to have the possibility of completion over a variable number of links. The need for operating in this fashion arises when calls may be completed directly to the called office or via one or more tandem or toll systems. Numbering difficulties of a plan which attempts to use tandems with direct dial control systems can be illustrated by reference to Fig. 5. Assume that A, B, C represent three direct dial control type offices in a 6-digit numbering plan area and that these are connected by direct trunks between offices. Office B is designated ACademy (22 on the dial) and office C is designated BLue Hills (25 on the dial). Analysis of the trunk layout in this network indicates, let us say, that trunking economies can be made by establishing a tandem and that the direct route from A to C is no longer economical as compared to the route via the proposed tandem. The digits 25 must now select a route via tandem. However, if we use both digits for selecting the route to tandem we have none left for selecting the route to office C at the tandem office. Since this plan will not work, let us see what results if we assume that the tandem trunks are selected by means of the first digit. Now all calls starting with the code digit 2 at office A must be routed via tandem and even though economies call for a direct route to the ACademy office from A we are forced to use the uneconomical route through tandem for this office. Actually we must consider the economy of routing the traffic for all offices whose codes begin with a given digit via tandem, or routing it over direct trunks, or we must change the designation of one of the offices. We could, of course, adopt the undesirable expedient of using non-universal numbering, i.e., numbering that varied by points of origin, as, for example, by introducing extra digits on calls through tandem from A to C and omitting them on calls from B to C.

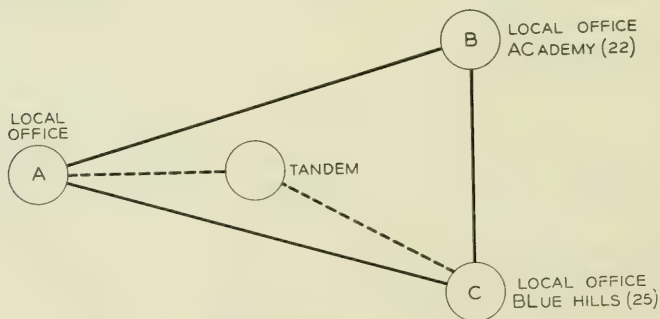


Fig. 5—Trunking scheme with a tandem office.

It is a situation such as has been described which has led to the practice, in some cases, of putting offices whose designations begin with the same first digit in the same building in step-by-step areas. This, of course, leads to restrictions.

Another alternative is to use selector repeaters. With these devices a "mitlaufer" action takes place in the local and tandem office selectors, i.e., both the local office selectors and the tandem office selectors follow the dial pulses until sufficient information is received to determine the route, whereupon the unneeded equipment is released. This equipment makes possible both the direct route to office B and the route via tandem to office C without an office designation change. However, selector repeaters are expensive and the cost of introducing them may be considerable. They also waste some trunk and equipment capacity because selector repeaters operate by seizing both local selectors and tandem trunks on every call. More often than not, perhaps, it would be cheaper to forego the trunk economy than to introduce the selector repeaters.

Now take the same network and assume common control equipment at all points. Prior to the introduction of the tandem the local offices translate the first two digits into information for selecting an outgoing trunk and then outpulse only the last four numerical digits directly to the called office. When the tandem is introduced, the translation at office A is changed to select a trunk to tandem on calls to BBlue Hills and to tell the sender at A to spill ahead the code digits or equivalent information as well as the line number for these calls. For calls to ACademy the existing arrangement is retained. There is no special problem at tandem since the code for the called office, BBlue Hills, is made available there. The translator at the tandem office tells the tandem sender to omit the office code digits in outpulsing to BBlue Hills.

There is an essential difference in the coding between direct dial control and common control which is obscured by the use of the same codes in the examples. In the direct dial control case the codes are route codes (sometimes called group codes); that is, the digits directly correspond to the route through the switches and are expended in the switching operations. In the common control case they are destination codes and it is not necessary to have them conform to the route nor are they used up in the switching process. Only common control systems can operate with destination codes. Therefore common control systems are required where it is necessary to route calls to some offices by direct trunks and calls to other offices via tandems without numbering restrictions.

Another example of a numbering difficulty with direct dial control systems tracing back to the use of route codes, is illustrated by an

extreme example in Fig. 6. This figure shows a multi-switch route through four automatic intertoll switching systems, A, B, C, D, to a customer whose listed number is 2345 in the central office, MAin 2. MAin 2 is in numbering plan area 217, a different area from that of the calling office. Typical digit combinations are shown at each place for reaching the next place with direct dial control systems. On a call from the A toll center area to the number MA 2-2345, the originating toll operator must dial 16 digits, such as 059 076 097 157 2345. Calls starting at intermediate points or in other networks use different numbers depending on the route. (Note that the route codes start with 0 or 1 to distinguish them from local codes.) It is rather obvious that dialing such combinations is cumbersome and requires elaborate routing information at each toll center. Intertoll calls through direct dial control systems are therefore generally limited to being switched at one place along the route, with infrequent use of two switching points.

However, with common control systems the situation is quite different. The originating point need dial only the ten digits of the destination 217 MA 2-2345. At each point except the one preceding the called area the full complement of digits is sent ahead. At that point the area code is dropped. At the last point, D, which is assumed to have direct circuits to the called office, MA 2 is skipped and 2345 is sent ahead. If calling and called points had been in the same numbering plan area, only seven digits would have been required. Note that since destination codes are used all points outside the numbering plan area dial the same 10 digits to reach a given line and all points within dial the same seven digits.

While only a small proportion of toll calls require multi-switch connections of the type just described, connections such as these are nevertheless required for an economically feasible nationwide network in which all calls are dialed to completion, and this objective cannot be attained practically without systems operating with destination codes.

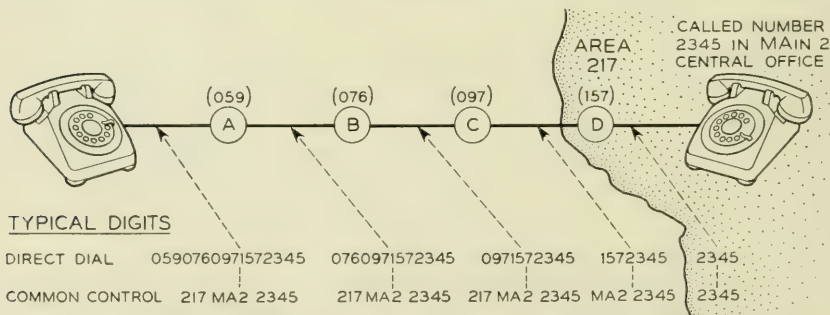


Fig. 6—Numbering with direct dial control and common control systems.

Also, as brought out later, destination codes are required in order to realize the important trunking economies of automatic alternate routing.

CODE CONVERSION

In passing, another feature of some common control systems, namely code conversion, can be brought out here because the illustration, Fig. 6, fits. Calls originating in a common control system can use office name codes (such as MA 2 for calls to the MA in 2 office) to reach destinations via step-by-step switching equipment where route codes (such as 157) are widely used. The translating equipment at the common control office can be arranged to substitute arbitrary digits for the office name code digits or in some cases to prefix arbitrary digits ahead of the called number. The arbitrary digits substituted or prefixed conform to the requirements of the office using route codes. In Fig. 6, office C when equipped with common controls could be arranged to convert MA 2 to 157, and therefore codes conforming to the nationwide numbering plan could be used for area 217 even though the calls were routed through step-by-step equipment.

RELATION BETWEEN TYPE OF SYSTEM AND TRUNKING ECONOMIES

The provision of a system which makes the most economical use of the trunk plant is important in any network but it is not as important in a small network as in a large one. Small networks can derive only small economies from arrangements which permit saving trunks. For example, in a single office network the trunks consist of wires running from originating to terminating equipment in the same building plus relatively cheap associated relay circuits. However, in a large toll network the trunks may include expensive repeaters, signaling equipments, carrier equipment and perhaps echo suppressors, as well as transmission channels running up to hundreds of miles in length and expensive toll relay circuits. For the larger networks there is therefore considerable urge to save as many trunks as possible. It is important therefore to operate these networks with switching plant that makes the most efficient use of the trunk plant by providing full access to groups, and to use an arrangement that permits the trunking economies of routes via tandems and of automatic alternate routing. These are features provided by common control systems and help explain why these systems are more attractive in the larger networks, both toll and local.

The cost of rearrangements for growth, new routes, load balancing and for restoring service under emergency conditions vary with the type

of system. Because of the flexibility of common controls such rearrangements are easier to make and usually cost less than in direct dial control systems. Also the frequency of rearrangements is greater in the larger places. Therefore this is another factor in favor of using common controls for those places.

SUPERIORITY OF COMMON CONTROL SYSTEMS WITH RESPECT TO SWITCH ACCESS

It has already been mentioned that the efficiency of trunks increases as the size of the group in which they are selected increases. Recognition of this fact early in the development of machine switching (about 1905) led to the invention of common controls. An ordinary step-by-step selector has access to only ten outlets on a level. Access to more than ten outlets can be obtained by providing graded multiple or by the use of rotary out-trunk switches,* or by combinations of these. Whenever it is necessary to employ graded multiple or rotary out-trunk switches, there is still some slight loss of efficiency as compared to full access.

In a system such as the panel system in which trunk hunting is a function of the selectors, the maximum number of trunks accessible to a call at any stage of selection is limited by the number of outlets accessible to the switch at that stage. A panel district or office selector, for example, can test a maximum of 90 trunks in a single group, 90 being the maximum number of terminals to which trunks can be assigned on a single panel bank, the remaining ten of the 100 terminals on a bank being reserved for overflow purposes. In the step-by-step system a corresponding limitation is avoided by a combination of graded multiple and rotary out-trunk switches with the penalty of a slight loss of efficiency. Marker systems avoid this limitation, also, by having the markers select trunks before they select the paths to the trunks. Crossbar systems with markers can readily test several hundred trunks for a given call. In some crossbar systems—No. 1, for example—trunks are tested in sub-groups of forty, therefore marker holding time is increased when there is more than one sub-group to be tested. This increase in marker holding time is largely avoided in systems like the toll crossbar systems by providing special testing arrangements in which a single indication per sub-group tells the marker which sub-group has one or more available trunks, whereupon the marker only tests the individual trunks of a sub-group in which it is assured that it can find an available trunk.

* A rotary out-trunk switch is arranged to hunt over a single group of outgoing trunks and to connect to an idle one. It is arranged for preselection and switches not in use will advance from busy trunks.

The maximum access of ten terminals on a level in ordinary step-by-step is not inherent in the system and might be overcome by a different switch design. A review of how a direct dial control system operates will help to clarify this point. At each switching stage, two actions take place. First, the switch follows the dial pulses until it reaches a group of outlets corresponding to the dialed digit. Then in the interval following this digit and before the pulses of the next digit arrive the switch hunts over the outlets for an idle path to reach the next stage. The number of paths from a switch level is therefore limited by the number of terminals the switch can hunt over in the interdigital interval. Assuming, for example, an interdigital interval of six-tenths of a second and a hunting speed of 100 terminals per second, 60 outlets could be provided. However, if such a high speed of hunting could be attained, and the 60 outlets were provided, 60 terminals would be required per group even for small ones which are in the majority. Hence such a switch would be wasteful of terminals. Direct dial control systems have generally employed switches with ten outlets per level although special arrangements such as twin levels have been employed to increase the number of outlets. A twin level switch provides terminals for two trunks at each rotary step and thus twenty trunks per level can be reached.

TRUNK ECONOMIES FROM TANDEM OPERATION WITH COMMON CONTROL SYSTEMS

An important factor in trunk economies is the ability to use tandems. The numbering difficulties that direct dial control systems have with tandems have already been discussed. Tandems permit major trunk economies on two scores. First, tandem routings take advantage of the efficiency which results from concentrating the smaller items of traffic and handling them over common trunk groups. Fig. 7 shows how this economy is attained. Ten offices completely interconnected by one-way trunks require 90 interoffice trunk groups. Ten offices interconnected only by way of tandem require only 20 groups. The groups by way of tandem are larger in size than the individual direct groups they replace and because of increased efficiency with group size fewer trunks are required.

There is a second possibility for an increase of efficiency, an example of which occurs when part of the offices are in business districts and part in residential districts. The peaks of trunked traffic from these different types of offices frequently occur at different hours, hence the trunks through tandem can be provided more economically for a given grade of

service than by an arrangement which must care for the peaks of each office separately. The non-coincidence of peaks of traffic of different types of offices permits economies both on trunks to tandem and trunks from tandem. For example, assume that a given office completes calls via tandem to some offices which have a morning busy hour and to others which have an evening busy hour. Then the group to tandem must provide capacity to handle the traffic for the busier hour of the two, but this capacity need care only for the peak traffic to part of the destinations. If individual direct groups had been provided instead of a common group to tandem, each group would have required capacity for its own peak, regardless of when it occurred. The common group to tandem therefore benefits by the noncoincidence of the peaks. A corresponding situation also occurs on trunks from tandem. Each group completes calls to a given destination from a number of originating offices whose peak hours may not coincide, and hence groups from tandem derive economies similar to those of the incoming groups to tandem.

Tandems are also required for alternate routing. Alternate routing is an arrangement to provide trunking economies by using a limited number of direct trunks for the traffic between two offices, and permitting the calls which do not find an available direct trunk to overflow to one or more tandems in succession. Because of the ability to load the direct circuits very heavily and yet provide good service by taking the overflow from and to a number of offices through a common tandem point, substantial economies are possible. Automatic alternate routing is practical only with common control systems. Common controls are needed to

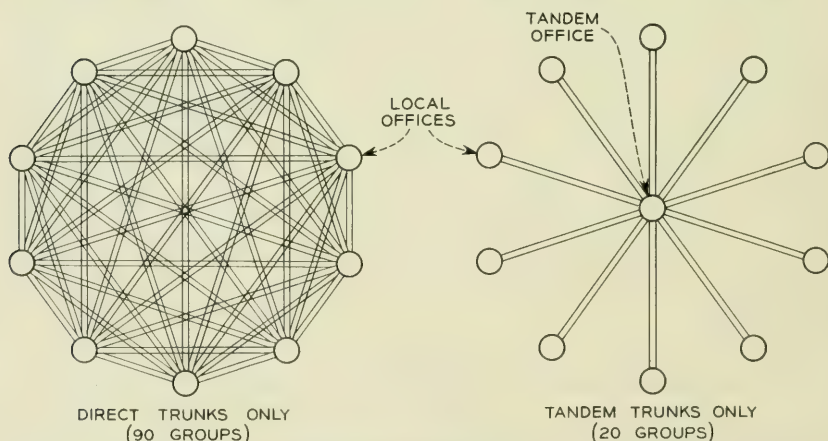


Fig. 7—Reduction of the number of trunk groups by the use of a tandem office.

provide the digit storage and digit spilling features in the office that does the alternate routing so that it can spill forward to the alternate route point the digits the latter requires.

Common controls have other advantages with respect to trunking which have already been covered in part. They also simplify the problems of assignment and load balancing as groups change in size or as new groups are added. An example of the difference in the methods of handling trunk growth in step-by-step and crossbar is of interest. In step-by-step when groups grow beyond 10 trunks a grade must be introduced in the switch wiring, or trunks must be sub-grouped or rotary out-trunk switches used. If further growth occurs, regrades must be made or rearrangements may be required in the sub-grouping or in the rotary out-trunk switches. In a crossbar system, however, in most cases added trunks are merely assigned to spare switch terminals which are left vacant for this purpose.

ROUTINGS FOR IRREGULAR CONDITIONS

Common controls are adapted to the efficient recognition and handling of irregular conditions such as permanent signals, vacant codes, and discontinued or temporarily intercepted lines.

Registers or senders detect line troubles which cause permanent signals or receivers off the hook by a timing circuit which waits for a short time for dialing to start. If the dialing does not start within the interval allowed the line is directed to a common group of permanent signal trunks which may appear before operators or at a test board. In No. 5 crossbar a trouble recorder card can be produced on which the location of the line in trouble is indicated. The step-by-step system indicates permanent signals by alarms to the maintenance force on a line group basis, and lines in trouble must be traced.

Vacant codes are detected by the translators, decoders and markers of common control systems and the calls are routed to a common trunk group which appears before operators or which returns "no such number tone." The corresponding arrangement in step-by-step requires connections from the switch multiple to operator or tone trunks.

In systems like No. 1 crossbar and No. 5 crossbar which have common controls in the terminating equipment, lines on which service has been discontinued or temporarily intercepted can be recognized by the markers and the calls rerouted to a common group of intercepting trunks. For example, temporary discontinuation of service is indicated by lifting a single cross-connection at the number group frame. In the step-by-step

system, however, one intercept trunk is commonly provided per 100 numbers and lines whose service is to be intercepted must be cross-connected to these trunks.

FURTHER ADVANTAGES OF COMMON CONTROL SYSTEMS ACCRUING FROM THEIR ABILITY TO OPERATE WITH TANDEMS

Some of the economies permitted by common control systems operating with tandems have been previously mentioned. Tandems are also useful because they provide centralized points at which special features can be concentrated with considerable saving.

For example, tandems are used for pulse conversion and for concentration of message charging equipment. Pulse conversion is needed when it becomes necessary to change from one type of pulsing to another, as, for example, on calls from a panel office to a step-by-step office. Panel can send out only revertive and panel call indicator pulses and step-by-step can receive only dial pulses. The two systems are therefore incompatible without special arrangements. The following are some of the plans which might be used for handling calls to step-by-step. First, all the panel senders could be modified to send out dial pulses. Second, spill senders could be provided at the outgoing trunks in the panel office or at the incoming trunks in the step-by-step office to receive, say, revertive pulses and convert them to dial pulses. Finally, if there is a tandem in the area, the tandem senders could be arranged (as they actually are) to accept revertive or panel call indicator pulses and send out dial pulses. The first two arrangements are usually more expensive than the last. Therefore, when pulse conversion is required it is generally done by routing calls via tandem.

To complete calls in the reverse direction, that is from step-by-step to panel, there is a requirement that is due to the use of the step-by-step system, namely that in cases where second dial tone is not employed the equipment at the called office or at an intervening tandem must be ready to accept the step-by-step pulses which are being dialed by the customer within a short time after the incoming trunk is seized. To meet this requirement, special high speed and costly link mechanisms are required to attach senders to incoming trunks or the incoming trunks must be arranged to record and store one or two digits. When calls are made between two systems both using senders, however, cheaper and slower link mechanisms can be employed because the calling senders are arranged to wait for a sender attached signal from the called office.

ADVANTAGES OF COMMON CONTROLS FOR AUTOMATIC RECORDING OF INFORMATION FOR CHARGING

The crossbar tandem system offers an economical method for making a record for charging purposes on multi-unit bulk billed calls called remote control of zone registration. At present this is limited to use with originating panel offices. The tandem is arranged to send back signals to the originating office for operating the customer's message register up to six times for the initial period on one call and also to operate it on overtime. Thus the application of extended customer dialing can be economically increased by applying this arrangement in places which cannot justify the registration arrangements available in the panel system itself which are economical only for a relatively heavy volume of this business. Local crossbar systems provide these features economically enough to obviate the need for tandem control of message registers for calls originating in the crossbar offices.

When tandem offices are required to control the equipment which records customers' charging data, they must be equipped with common controls if the arrangement is to be economical. The data includes the origin of the call—the particular trunk group incoming to tandem over which the call arrives—and the destination—the called office code. These elements must be analyzed and combined to determine the basis for the amount charged. Since elaborate equipment is required for these functions, economy must be attained by providing a minimum amount of equipment to do the job. This objective is accomplished by providing the required features in the common controls. In tandems arranged for remote control of zone registration, for example, the number of times the customer's message register is operated is determined partly by the choice of trunk group at the originating office and partly by the tandem markers.

In addition to remote control of zone registration, there are several other methods of determining and recording charging data which also require the use of common control equipment. These are automatic ticketing, automatic message accounting and coin zone dialing.

In automatic ticketing, which is used with step-by-step systems, calls which are to be ticketed are directed to outgoing trunks which select senders and other common equipment which determine the calling line number, reconstruct the called office code and store and output the digits required for selections beyond the local office. The calling line number and the called office code are transmitted by the common equipment to the outgoing trunk which is equipped with a ticket printing

device which prints this information and other data required for charging. The tickets can be used for bulk bills as well as detail records since they can be summarized at the accounting center by manual methods for calls on which detail information is not required.

Automatic message accounting is used with crossbar systems both for bulk billing and detailed call records. With this system the data required for charging is perforated on paper tape by common central office equipment. The arrangement has been described in the technical literature* and will not be further described here.

Both the ticketing method and automatic message accounting require the collection of a large amount of data and the ability to do a complicated job in handling and recording this data. This demands elaborate and expensive equipment which is practical only when provided on a common basis so that it can be called into service for a short time and then restored to the common pool for other calls.

Direct dial control systems without common controls can only have message registers on the line and therefore can handle nothing but bulk billed calls. Furthermore because of the expense of arrangements for determining multiple unit charge data and for operating the message register more than once on a call, multiple operation of message registers on individual calls is not practical.

From coin stations in direct dial control systems the customer may dial calls only to offices within the local charge zone. However, in panel and crossbar areas the "coin zone dialing" arrangement is available to permit coin customers to dial beyond the local zone. With this plan calls are routed to a tandem office where completion is delayed until an operator can plug into the trunk to tandem and supervise the collection of the required coins. The amount to be collected is indicated by trunk lamps which appear in a switchboard multiple. Common controls enter into this scheme at the originating office to route the call to tandem and to determine the charge, and at the tandem office so that the digits can be stored while the call is held up prior to collection of the coins.

TYPES OF PULSING

Direct dial control systems are restricted to operation with dial pulses and are usually limited to pulsing speeds of about 10 pulses per second and about one digit per second. Dial pulsing has range limitations which can be overcome by the addition of pulse repeaters at appropriate points.

Common control systems store the digits in senders which can regen-

* *A.I.E.E. Transactions*, **69**, Part 1, pp. 255 to 268, 1950.

erate them in various types and combinations of types of pulsing. Types of outpulsing found today in various systems include revertive, panel call indicator, dial pulsing, dc key pulsing, and multi-frequency pulsing. Panel sender tandem and No. 4 toll can also send digital information ahead to operators by the call announcer method which uses voice announcements derived from recordings on film. Provision for receiving and sending several types of pulsing in one system makes it more flexible since it can then connect to a variety of equipments. Regenerating the pulses adds to the range without the need of adding pulse repeaters.

Some of the advantages which common control systems derive from the ability to operate with a modern type of pulsing can be brought out by a brief description of multi-frequency pulsing which is a relatively recent development. Digital information is transmitted over any facility capable of handling voice by sending spurts of alternating current which consist of pairs of frequencies in the voice range selected out of five frequencies. There are ten such pairs. At the receiving end a check is made to insure that exactly two frequencies are received for each digit. When only one or more than two frequencies per digit are detected the call is not set up but a reorder signal is returned to the originating end. In addition to the advantages of being capable of transmission over voice facilities, including repeaters and carrier systems, and of providing checks for accuracy, this type of pulsing can be transmitted at the rate of seven digits per second at present. Operators can be provided with keysets capable of sending MF pulses into either local or distant switching equipment with improved operating resulting from the higher speed and other advantages of MF pulsing.

It is quite feasible to add new types of pulsing to common control systems. Multi-frequency pulsing has only recently been added to crossbar tandem, for example, although it has been in use with other crossbar systems for some time. In this case it required the development of new senders capable of receiving and sending the MF pulses. The addition of these senders, even in existing offices, is not a difficult job.

IMPROVED STATION APPARATUS

The stations in most exchanges are provided with dials which operate at approximately 10 pulses per second. In step-by-step exchanges this pulsing speed is the maximum permitted by the capabilities of the switches. In panel and crossbar areas the common equipment is capable of operating with higher speed dial pulsing, and PBX and central office operators in these areas are usually given dials that operate at about 18 pulses per second.

Even fast dials are inefficient as compared to the push button keysets used by operators for key pulsing and it is obvious that subscriber sets with push buttons would be faster and more convenient than dials. Such sets were used at Media, Pa., on an experimental basis and have functioned in a highly satisfactory manner. Their introduction merely required the design and installation of registers to receive the pulses they generate. This was done with little difficulty or expense at the central office end. However, with ordinary step-by-step systems such devices are impractical because of the short interdigital interval they allow and because of the cost of adding the pulse receiving equipment in every selector and of providing translation to change the key pulses into a form to drive the switch.

CLASSES OF SERVICE

Differences in the handling of calls from non-coin, coin and PBX lines and differences in rate treatments require the recognition of classes of customers at the central office. In step-by-step separate groups of line finders are provided to permit segregation in classes and where routings for different classes vary, separate selector multiples are required for these routings. Class distinctions within a line finder group can be made by normal post springs and by marking a fourth conductor in the line circuit.

Common control systems permit the economical handling of many classes of service. The No. 5 crossbar, for example, is most flexible in this respect. As many as thirty classes of service can be handled in a single line link frame, including coin and non-coin. Special handling, reroutes and restrictions are mostly functions of the common controls and inefficiencies due to segregation of traffic in small groups of switching equipment are largely avoided.

DOUBLE CONNECTIONS

In systems such as panel and step-by-step in which selectors do the hunting, several selectors may be hunting over the same terminals simultaneously, and since there is an unguarded interval just after an idle terminal has been found before it is made busy by the release of the busy testing relay, double connections occur. Considerable effort and expense have been expended to reduce the probability of double connections in these systems. In systems which employ markers, on the other hand, the trunk testing schemes do not normally permit double connections to

occur. In most marker systems a lockout arrangement permits only one marker at a time to test trunks in a given group. There are cases where trunks are common to two offices and two markers are allowed to test trunks simultaneously. In these cases special circuit arrangements are provided at nominal expense to avoid double connections. Modern common control systems with markers are, therefore, free of double connections resulting from weaknesses of the system and they can occur only as a consequence of defects in circuits or apparatus.

THEORETICAL OFFICES

It is sometimes desirable to assign more than one office designation to customers in a single central office unit. A new unit may be planned for sometime in the future and if growth on the existing unit can be taken with a new office designation, then when this new office is placed in service it can be done without directory changes by transferring a block of lines from the old unit. Another occasion for assigning more than one designation to a single unit arises when customers served by the unit are in two rate zones, and service to lines in one of the rate zones must be restricted or extra charges collected. The lines served by an additional designation are called a theoretical office. Common control systems handle theoretical offices with little difficulty. In the first case mentioned the translating equipment in the originating offices recognizes that the physical office and theoretical office designations require identical treatment until the new unit is cut into service at which time translator cross-connection changes take care of the new routings. Where different rate treatments are involved, records for billing purposes depending on both the origin and destination of the call can be made by methods previously mentioned. In some cases where the billing data is determined at a tandem office and different treatments for the same destinations must be given to customers calling from one office, split trunk groups must be provided to tandem, one for each treatment.

In the step-by-step system, theoretical offices can be opened up by multiplying two selector levels together. For example, if the physical office is designated 25 and it is desired to open a theoretical office, say 26, the 5 and 6 levels on the proper second selectors in the network can be strapped until the 26 office is changed to a physical office. At that time the levels are split and trunks to the new office are connected to the 6 levels of the second selectors. Restrictions in reaching blocks of numbers can be applied by splitting selector multiples and intercepting calls to restricted blocks from one of the splits.

ADAPTABILITY TO NEW SERVICE FEATURES

One of the major advantages of common controls, which has been covered in part but which deserves further emphasis, is their adaptability to new service features. Key sets and new dialing devices can be introduced at customers' stations and operator positions by readily feasible modifications of registers and senders. New pulsing schemes can also be introduced as they are developed as evidenced by the introduction of multi-frequency pulsing over the past few years. Nationwide customer dialing, now under development, can be readily introduced in existing common control systems by economical modifications without the use of either directing codes or second dial tone. Step-by-step systems require at least partial senderization to provide equivalent service. In short, the flexibility of common controls and the concentration of the control elements in a relatively few circuits makes the addition of new service features easier and more economical than in direct dial systems.

MAINTENANCE ASPECTS

Experience has shown that switches with a large amount of motion, especially those with brushes which wipe over bank terminals, tend to wear excessively and require considerable maintenance effort and even replacement, at times. On the other hand, switches with short motions and relay-like action require little maintenance and tend to have long life. Furthermore, the switches which employ wiping brushes mostly use base metal contacts, whereas relay-like switches can readily be equipped with precious metal contacts—and in most cases are so equipped—with the elimination of the transmission noise to which base metal contacts are subject. The crossbar switch is a relay type of switch with precious metal contacts and considerations such as those mentioned influenced its adoption. The advantages of relay type switches are not necessarily limited to common control systems since such switches have been used in direct dial control systems. The first use of the crossbar switch in Sweden was in a step-by-step system, for example. However, economical arrangements for using such switches in large systems require markers. This is because economy must be achieved by having more than one call occupy a switch at a time and marker control is necessary to attain this.

Important maintenance advantages have been introduced in systems using decoders and markers. In this category are the self-checking features, second trials with changed order of preference, and trouble reporting features. In No. 5 crossbar the ability to report the location of a line

with a permanent signal by perforating a trouble recorder card has eliminated the need for tracing permanents.

A number of schemes are employed to detect troubles in markers and decoders and in circuits which connect to them. These include detectors for wrong sequences of operations, wrong combinations of relays, excessive current, false potential and lack of continuity. These are generally introduced at small cost since the circuits to which they are applied are small multipliers. However, some of them do a major job of testing since they reach out and test the numerous elements of the switching system to which markers have access. In this category are the tests of the cross-bar linkages for opens, false grounds and double connections, tests of the switch crosspoints for continuity, tests of lines for false grounds, and for receivers off the hook on coin first coin lines.

To obtain clear trouble records, markers are designed with interlocked progress signals. This has made trouble analysis easier and has tended to improve design by eliminating relay races.

Starting with the panel system tests have also been introduced in senders for detecting open and reversed trunks. These tests have been of considerable help in maintaining outside plant and in detecting conditions that could lead to false charges.

DISADVANTAGES OF COMMON CONTROLS

Up to this point the stress has been mainly on the advantages of common controls. There are also some disadvantages. One of the major ones is the substantial getting started cost due to the necessity of providing a minimum amount of common equipment. This minimum is provided to maintain operation in case of trouble and during intervals when, for example, cross-connections require change because of changed or added routes. The minimum requirements establish economic barriers which tend to prohibit the economical use of common controls for small isolated systems.

Another disadvantage is the performance of common control systems under severe and protracted overloads. Experience with these systems indicates that although they compare quite favorably to direct dial control systems with respect to capability of handling moderate overloads, they are not able to handle severe overloads as well. In part this is a consequence of the fact that elements in common control systems are used at high efficiency and hence there is relatively less free equipment at full load for soaking up an overload than there is in systems that operate with smaller and less efficient groupings. Whenever the number

of calls presented to the system exceeds the capacity of the common control elements provided, the excess calls are delayed. The things which customers, operators and connecting switching machines do when they encounter delays tend to aggravate the overload. The reactions of operators and customers to delays can be illustrated by two examples.

The first is taken from the operation of a network of No. 4 toll crossbar systems when one of the No. 4's is heavily overloaded. Operators placing calls through the overloaded system encounter, let us say, an abnormal number of "no circuit" conditions in the outgoing trunks. This causes them to make additional attempts to get circuits. These additional attempts plus the excessive number of first attempts overload the markers. Sender holding time is then increased because of delays in connecting to the markers and this, added to the abnormal number of sender usages, results in a further shortage of senders. Operators trying to place calls through the system are therefore slowed down because of slow "sender attached" signals. (These are the signals which tell the operators that they can start keying or dialing.) Senders in connecting systems are also delayed waiting for senders to become idle in the overloaded office. The overload therefore tends to spread to all connecting systems.

However, it is possible to provide remedies which limit the reaction to the overloaded system. These remedies are arrangements to rapidly clear out senders waiting for senders ahead. Automatic alternate routing is also useful in routing traffic around overloaded systems.

The second example is taken from local systems. Here the reaction of customers to delays compounds the overload. A severe overload results in a shortage of senders, much as described above. A shortage of senders in a local system causes dial tone delays. There are always some customers who either do not listen for dial tone or who will not wait very long for it, and who start to dial before senders are attached to their lines. The result of such dialing is either a partial digits condition under which the sender waits for a considerable interval for a full complement of digits, or a wrong number when the first digit is clipped. The delays reduce sender capacity still further and the wrong numbers further increase the attempts. The load "snowballs" and the ability of the system to handle calls degenerates.

Here again arrangements are available to control the overload. These include features for blocking calls before they reach the senders and markers, and for returning paths busy signals with a minimum of common circuit holding time.

While there is, then, a somewhat greater capacity for overloads in step-by-step because of less efficient use of equipment, common control

systems do a good job of handling moderate overloads and, by provision of load control features, can operate satisfactorily even with severe overloads.

From a maintenance standpoint, a disadvantage of common controls is the relative complexity of the circuits. While this has introduced a training problem, maintenance forces have had no difficulty in acquiring the knowledge needed to do a competent maintenance job.

CONCLUSION

The full fledged common control systems exemplified by the crossbar local and toll systems have a number of important advantages over systems where the switches are driven directly by the customer's dial. The advantages arise largely from the ability to store digits, to translate them, use them flexibly for switching within the office, and transmit as many of them as desired to distant points for subsequent switching operations. The digits can be converted to others of different value whenever it is advantageous to do so. The inherent flexibility of common control equipment makes it possible to adopt any kind of numbering plan for a local area or a nationwide network that is best suited for the purpose without regard to the manner in which calls will be trunked from one point to another. Codes can be assigned at will to represent destinations and the best route for the call can always be taken. The best route may in some cases involve tandem operation or even a half-dozen switches in tandem. It may be the route selected as an alternate after previous trial of one or more other routes. A connection may be set up between offices of different types and over trunk groups requiring different forms of pulsing. These conditions may be met by common control equipment and the ability to meet such conditions makes it possible to provide cheap step-by-step equipment in places for which it is best suited, compensating for some of its deficiencies with common control equipment in other places.

With marker type common controls, trunk groups out of an office can be of any desired size regardless of the switch design. The individual crossbar switch, for example, gives access to only ten or twenty outlets as normally wired but full access single trunk groups of hundreds of trunks can be employed in some crossbar systems.

Schemes for recording billing data, aside from the relatively simple ones where metering equipment is associated with the customer's line and operated once per call, make use of common control equipment. This seems to be necessary where detail records must be made on individual calls for charging purposes.

As improvements in the art are made they can more readily be incorporated in common control systems than in step-by-step systems. For example, new subsets which may employ keys or other sending devices different from the dial can be accommodated by provision of proper facilities in senders and registers. Also, improved high speed pulsing arrangements can be easily incorporated in systems which do not require the switches themselves to be directly driven by pulses from the calling device.

BIBLIOGRAPHY

1. Bailey, W. J., "Lorimer automatic exchange at Hereford," *P. O. Elect. Engrs. J.*, **6**, Part 2, pp. 97-155, July, 1913.
2. Aitken, William, *Automatic telephone systems*, D. Van Nostrand Co., N. Y., 1924.
3. Miller, K. B., *Telephone theory and practice*, McGraw-Hill, N. Y., 1933.
4. Rorty, M. C., "How the theory of probability may be applied to telephone traffic," *Western Electrician*, **36**, p. 356, May 6, 1905, Abstract.
5. Craft, E. B., and others, "Machine switching telephone system for large metropolitan areas," *Bell System Tech. J.*, **9**, pp. 266-274, Jan., 1934.
6. Bronson, F. M., "Tandem operation in the Bell System," *Bell System Tech. J.*, **15**, pp. 380-404, 1936.
7. Scudder, F. J., and Reynolds, J. H., "Crossbar dial telephone switching system," *Bell System Tech. J.*, **18**, pp. 76-118, Jan., 1939.
8. Abraham, L. G., and others, "Crossbar toll switching system," *A.I.E.E. Trans.*, **63**, pp. 302-309, 1944.
9. Friend, O. A., "Automatic ticketing of telephone calls," *A.I.E.E. Trans.*, **63**, pp. 81-88, 1944.
10. Smith, A. B., "The 'director' for automatic telephone switching systems," *A.I.E.E. Trans.*, **67**, pp. 611-619, 1948.
11. Meszar, J., "Fundamentals of the AMA system," *A.I.E.E. Trans.*, **67**, Part 1, pp. 255-269, 1950.

Mathematical Theory of Laminated Transmission Lines—Part II

By SAMUEL P. MORGAN, JR.

This part of the paper continues the analysis of the low-loss, broad-band, laminated transmission lines proposed by A. M. Clogston, and deals particularly with "Clogston 2" lines, in which the entire propagation space is filled with laminated material.

TABLE OF CONTENTS

	Page
VIII. Principal Mode in Clogston 2 Lines with Infinitesimally Thin Laminae	1121
IX. Partially Filled Clogston Lines. Optimum Proportions for Principal Mode	1133
X. Higher Modes in Clogston Lines	1150
XI. Effect of Finite Lamina Thickness. Frequency Dependence of Attenuation in Clogston 2 Lines	1163
XII. Effect of Nonuniformity of Laminated Medium	1181
XIII. Dielectric and Magnetic Losses in Clogston 2 Lines	1201
Appendix II: Optimum Proportions for Heavily Loaded Clogston Cables	1203
Appendix III: Power Dissipation in a Hollow Conducting Cylinder	1204

VIII. PRINCIPAL MODE IN CLOGSTON 2 LINES WITH INFINITESIMALLY THIN LAMINAE

In Part I* of this paper we have set up a general mathematical framework for the analysis of Clogston-type laminated transmission lines and have applied it to Clogston 1 lines having laminated conductors, but with the total thickness of the laminations small compared to the overall dimensions of the line, so that most of the forward power flow takes place in the main dielectric. In Part II we shall consider Clogston 2 lines, which instead of containing a main dielectric have the propagation space entirely filled with laminations; and we shall also derive results, in Sections IX and X, for the general laminated transmission

* S. P. Morgan, Jr., *Bell System Tech. J.*, **31**, 883 (1952). Since the two parts of the paper are very closely related, the sections, equations, figures, and footnotes have been numbered consecutively throughout the whole paper. A table of symbols appears at the end of Part I.

line in which the relative fractions of space occupied by the main dielectric and the laminations are arbitrary.

A parallel-plane Clogston 2 line is shown schematically in Fig. 10. It consists of a stack of alternate layers of conducting and insulating material, whose total thickness is a . As before, the electrical constants of the conducting and insulating layers are denoted by μ_1 , g_1 and ϵ_2 , μ_2 respectively; and the fraction of conducting material in the stack is called θ . The stack is bounded at $y = \pm \frac{1}{2}a$ by sheaths whose normal surface impedance is $Z_n(\gamma)$, where γ is the longitudinal propagation constant of the mode under consideration.

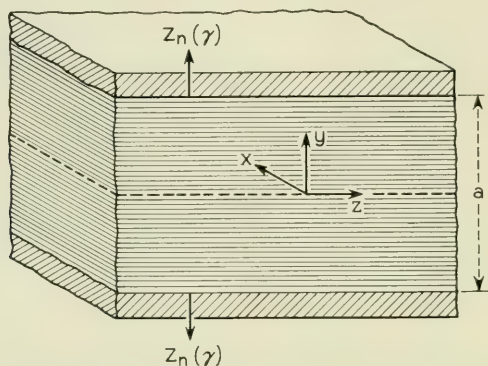


Fig. 10—Parallel-plane Clogston 2 transmission line.

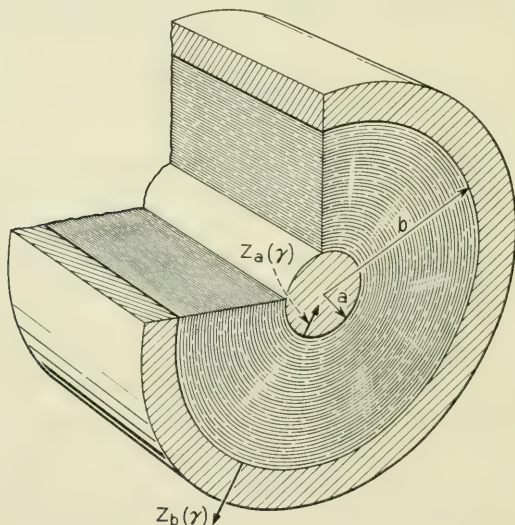


Fig. 11—Coaxial Clogston 2 transmission line.

The cross section of a coaxial Clogston 2 cable is shown schematically in Fig. 11. It consists of a laminated coaxial stack bounded internally by a cylindrical core of radius a , which may be equal to zero so far as the theoretical analysis is concerned, and externally by a cylindrical sheath of radius b . We denote the radial impedance looking into the core at $\rho = a$ by $Z_a(\gamma)$, and the radial impedance looking into the sheath at $\rho = b$ by $Z_b(\gamma)$.

In this section we shall assume the laminae to be infinitesimally thin, so that the stack may be regarded as a homogeneous, anisotropic medium, completely characterized by its average electrical constants. The case of finite lamina thickness will be treated in Section XI. We shall neglect dielectric and magnetic dissipation throughout, except in Section XIII.

For modes of the type which we consider, whose only field components are H_x, E_y, E_z in the plane line or H_ϕ, E_ρ, E_z in the coaxial line, the average electrical constants of the stack are given by equations (90) of Section III, namely,

$$\begin{aligned}\bar{\epsilon} &= \epsilon_0/(1 - \theta), \\ \bar{\mu} &= \theta\mu_1 + (1 - \theta)\mu_2, \\ \bar{g} &= \theta g_1.\end{aligned}\tag{268}$$

As observed in Section III, Maxwell's equations for the average fields in such an artificial anisotropic medium take the form, for a plane stack,

$$\begin{aligned}\partial \bar{H}_x / \partial z &= i\omega \bar{\epsilon} \bar{E}_y, \\ \partial \bar{H}_x / \partial y &= -\bar{g} \bar{E}_z, \\ \partial \bar{E}_y / \partial z - \partial \bar{E}_z / \partial y &= i\omega \bar{\mu} \bar{H}_x;\end{aligned}\tag{269}$$

while for a cylindrical stack,

$$\begin{aligned}\partial \bar{H}_\phi / \partial z &= -i\omega \bar{\epsilon} \bar{E}_\rho, \\ \partial(\rho \bar{H}_\phi) / \partial \rho &= \bar{g} \rho \bar{E}_z, \\ \partial \bar{E}_z / \partial \rho - \partial \bar{E}_\rho / \partial z &= i\omega \bar{\mu} \bar{H}_\phi.\end{aligned}\tag{270}$$

We wish to determine the modes which can propagate in the laminated medium when guided by plane or cylindrical impedance sheets. This problem was solved for a homogeneous, isotropic dielectric in Section II of Part I; and the method of solution is so similar for the anisotropic

medium that we shall omit details of the analysis and pass at once to the results.

In the parallel-plane line, the modes for which \bar{H}_x is an even function of y about the center plane $y = 0$ have field components given by

$$\begin{aligned}\bar{H}_x &= \text{ch } \Gamma_t y \, e^{-\gamma z}, \\ \bar{E}_y &= -\frac{\gamma}{i\omega\bar{\epsilon}} \text{ch } \Gamma_t y \, e^{-\gamma z}, \\ \bar{E}_z &= -K \text{sh } \Gamma_t y \, e^{-\gamma z},\end{aligned}\tag{271}$$

up to an arbitrary amplitude factor, where Γ_t and K are defined, as in Section III, by

$$\Gamma_t = \left[\frac{i\bar{g}}{\omega\bar{\epsilon}} (\omega^2 \bar{\mu}\bar{\epsilon} + \gamma^2) \right]^{\frac{1}{2}},\tag{272}$$

$$K = \Gamma_t / \bar{g} = \left[\frac{i}{\omega\bar{\epsilon}\bar{g}} (\omega^2 \bar{\mu}\bar{\epsilon} + \gamma^2) \right]^{\frac{1}{2}}.\tag{273}$$

Matching impedances at the boundaries $y = \pm \frac{1}{2}a$ leads to the condition

$$K \tanh \frac{1}{2} \Gamma_t a = -Z_n(\gamma).\tag{274}$$

In the odd case, the fields are given by

$$\begin{aligned}\bar{H}_x &= \text{sh } \Gamma_t y \, e^{-\gamma z}, \\ \bar{E}_y &= -\frac{\gamma}{i\omega\bar{\epsilon}} \text{sh } \Gamma_t y \, e^{-\gamma z}, \\ \bar{E}_z &= -K \text{ch } \Gamma_t y \, e^{-\gamma z},\end{aligned}\tag{275}$$

up to an arbitrary amplitude factor, and the boundary condition becomes

$$K \coth \frac{1}{2} \Gamma_t a = -Z_n(\gamma).\tag{276}$$

General expressions for the field components in the coaxial line are

$$\begin{aligned}\bar{H}_\phi &= [A I_1(\Gamma_t \rho) + B K_1(\Gamma_t \rho)] e^{-\gamma z}, \\ \bar{E}_\rho &= \frac{\gamma}{i\omega\bar{\epsilon}} [A I_1(\Gamma_t \rho) + B K_1(\Gamma_t \rho)] e^{-\gamma z}, \\ \bar{E}_z &= K [A I_0(\Gamma_t \rho) - B K_0(\Gamma_t \rho)] e^{-\gamma z},\end{aligned}\tag{277}$$

where A and B are arbitrary constants and Γ_l and K are defined as before. The boundary conditions at $\rho = a$ and $\rho = b$ take the form

$$\begin{aligned} K \frac{AI_0(\Gamma_l a) - BK_0(\Gamma_l a)}{AI_1(\Gamma_l a) + BK_1(\Gamma_l a)} &= Z_a(\gamma), \\ K \frac{AI_0(\Gamma_l b) - BK_0(\Gamma_l b)}{AI_1(\Gamma_l b) + BK_1(\Gamma_l b)} &= -Z_b(\gamma), \end{aligned} \quad (278)$$

and these equations can be satisfied by values of A and B that are not both zero if and only if

$$\frac{KK_0(\Gamma_l a) + Z_a(\gamma)K_1(\Gamma_l a)}{KI_0(\Gamma_l a) - Z_a(\gamma)I_1(\Gamma_l a)} = \frac{KK_0(\Gamma_l b) - Z_b(\gamma)K_1(\Gamma_l b)}{KI_0(\Gamma_l b) + Z_b(\gamma)I_1(\Gamma_l b)}. \quad (279)$$

Now K is given in terms of Γ_l by equation (273), while from equation (272) we have

$$\gamma^2 = -\omega^2 \bar{\mu} \bar{\epsilon} - (i\omega \bar{\epsilon} / \bar{g}) \Gamma_l^2. \quad (280)$$

Hence if the dependence of the boundary impedances on γ is known, equations (274) and (276) for the plane line and equation (279) for the coaxial line are transcendental relations from which in principle we may determine Γ_l , and therefore γ , for each mode of the type that we are considering. If the value of Γ_l for a particular mode satisfies the inequality

$$\frac{1}{8} \left| \frac{\Gamma_l^2}{\omega \bar{\mu} \bar{g}} \right|^2 \ll 1, \quad (281)$$

then on taking the square root of the right side of equation (280) by the binomial theorem, we find that the attenuation and phase constants of the given mode are approximately

$$\alpha = \text{Re } \gamma = -\text{Re } \frac{\Gamma_l^2}{2\bar{g}\sqrt{\bar{\mu}/\bar{\epsilon}}}, \quad (282)$$

$$\beta = \text{Im } \gamma = \omega \sqrt{\bar{\mu} \bar{\epsilon}} - \text{Im } \frac{\Gamma_l^2}{2\bar{g}\sqrt{\bar{\mu}/\bar{\epsilon}}}. \quad (283)$$

Throughout the rest of this section we shall consider only the lowest or principal mode. In a parallel-plane line the principal mode corresponds to the lowest root in Γ_l (that is, the root having the smallest modulus) of equation (274), which may be written in the form

$$\frac{1}{2} \Gamma_l a \tanh \frac{1}{2} \Gamma_l a = -\frac{1}{2} \bar{g} a Z_n(\gamma). \quad (284)$$

We may express γ in terms of Γ_t by equation (280), and so if $Z_n(\gamma)$ varies with γ in any reasonably simple way, or better yet if $Z_n(\gamma)$ is essentially independent of γ in the range of interest, equation (284) may be solved numerically for Γ_t by successive approximations.

A numerical solution of equation (284) is, however, rarely necessary, since the right-hand side of the equation is just the ratio of the sheath impedance $Z_n(\gamma)$ to the resistance "per square", namely $1/(\frac{1}{2}\bar{g}a)$, of all the conducting layers in a stack of thickness $\frac{1}{2}a$ in parallel, and this ratio will almost always be large compared to unity. This is another way of saying that the total one-way conduction current in the stack is large compared to the sum of the conduction and displacement currents in either sheath. Even if the sheaths are infinitely thick metal plates of conductivity g_1 , we have from equation (79) of Section III, since $\gamma \approx i\omega\sqrt{\bar{\mu}\bar{\epsilon}}$,

$$\frac{1}{2}\bar{g}aZ_n(\gamma) = \frac{1}{2}\theta g_1 a \eta_1 = (1 + i)\theta a / 2\delta_1, \quad (285)$$

and for most frequencies of interest the thickness $\frac{1}{2}\theta a$ of conducting material in half the stack will be several times the skin thickness δ_1 in the metal. If the medium outside the stack is free space, then $Z_n(\gamma)$ will be a few hundred ohms and a fortiori the right side of (284) will be large compared to unity. So long as the inequality

$$|\frac{1}{2}\bar{g}aZ_n(\gamma)| \gg 1 \quad (286)$$

is satisfied, the lowest root of (284) will be approximately

$$\Gamma_t = i\pi/a; \quad (287)$$

and so from (282) and (283) the attenuation and phase constants of a plane Clogston 2 line with infinitesimally thin laminae and high-impedance walls are

$$\alpha = \frac{\pi^2}{2\sqrt{\bar{\mu}/\bar{\epsilon}} \bar{g}a^2}, \quad (288)$$

$$\beta = \omega\sqrt{\bar{\mu}\bar{\epsilon}}. \quad (289)$$

To this approximation, there is neither amplitude nor phase distortion.

The principal mode in a coaxial Clogston 2 corresponds to the lowest root in Γ_t of equation (279). To solve this equation numerically with finite boundary impedances $Z_a(\gamma)$ and $Z_b(\gamma)$, while possible in principle, would evidently be a major undertaking. We shall therefore assume throughout the present paper that the total conduction and displacement currents flowing in the core and the sheath are negligible compared

to the conduction currents in the laminated medium. This is equivalent to assuming that the boundary impedances $Z_a(\gamma)$ and $Z_b(\gamma)$ are effectively infinite, so that equation (279) reduces to the simple form

$$\frac{K_1(\Gamma_\ell a)}{I_1(\Gamma_\ell a)} = \frac{K_1(\Gamma_\ell b)}{I_1(\Gamma_\ell b)}. \quad (290)$$

Equation (290) may be converted to one involving ordinary Bessel and Neumann functions by the substitution

$$\Gamma_\ell^2 = -\chi^2, \quad \Gamma_\ell = i\chi. \quad (291)$$

Then since

$$\begin{aligned} I_n(\Gamma_\ell \rho) &= i^n J_n(\chi \rho), \\ K_n(\Gamma_\ell \rho) &= \frac{1}{2} \pi i^{-(n+1)} [J_n(\chi \rho) - i N_n(\chi \rho)], \end{aligned} \quad (292)$$

the equation may easily be transformed into

$$J_1(\chi a) N_1(\chi b) - J_1(\chi b) N_1(\chi a) = 0. \quad (293)$$

For any given value of the ratio a/b , equation (293) has an infinite number of real roots in χ . The lowest root χ_1 has been tabulated¹⁸ as a function of b/a , and may be written in the form

$$\chi_1 = \frac{\pi f_1(a/b)}{b - a}, \quad (294)$$

where $f_1(a/b)$ is a monotone decreasing function of a/b which is equal to 1.2197 when $a/b = 0$ and to 1 when $a/b = 1$. Hence the attenuation and phase constants of the principal mode in a coaxial Clogston 2 with infinitesimally thin laminae and high-impedance walls are

$$\alpha = \frac{\pi^2 f_1^2(a/b)}{2\sqrt{\mu/\epsilon} \bar{g}(b - a)^2}, \quad (295)$$

$$\beta = \omega \sqrt{\mu \epsilon}, \quad (296)$$

and again to this approximation there is neither amplitude nor phase distortion.

Comparing equations (288) and (295), we see that the attenuation constant of the principal mode in a coaxial Clogston 2 with infinitesimally thin laminae (that is, the low-frequency attenuation constant if the laminae are of finite thickness) is equal to the attenuation constant of

¹⁸ E. Jahnke and F. Emde, *Tables of Functions*, fourth ed., Dover, New York, 1945, pp. 204-207. What we call $\pi f_1(a/b)$ is tabulated by Jahnke and Emde, p. 205, as $(k - 1)x_1^{(1)}$, where $k = b/a$, while our $f_1(a/b)$ is plotted as $1 + \alpha$ on p. 207.

the principal mode in a plane Clogston 2 times the factor $f_1^2(a/b)$, provided that the thickness of the plane stack is equal to the thickness $b - a$ of the coaxial stack. The functions $f_1^2(a/b)$ and $f_1^2(a/b)/(1 - a/b)^2$ are plotted against a/b in Fig. 12. From the plots it is apparent that $f_1^2(a/b)$ decreases steadily from a value of 1.488 at $a/b = 0$ to 1 at $a/b = 1$, while $f_1^2(a/b)/(1 - a/b)^2$ increases steadily from 1.488 at $a/b = 0$ to infinity at $a/b = 1$. Therefore if the stack thickness $b - a$ is fixed, the attenuation constant will be smaller the greater is the mean radius of the stack; while if the outer radius b is fixed, the attenuation constant will be reduced by reducing the radius a of the inner core, and the lowest attenuation will be achieved when $a = 0$.

It should be noted that our expressions for the attenuation and phase constants of Clogston 2 lines cannot be valid down to the mathematical limit of zero frequency, since the inequality (281), on which we based the approximations (282) and (283) for α and β , will ultimately break down as the frequency approaches zero. A similar failure of the approximate expressions which we used for the attenuation and phase constants of Clogston 1 lines was pointed out in Section II of Part I. Here, as before, we shall limit the use of the term "low frequency" to frequencies still high enough so that the attenuation per radian is small and the approximate formulas (282) and (283) for α and β are valid.

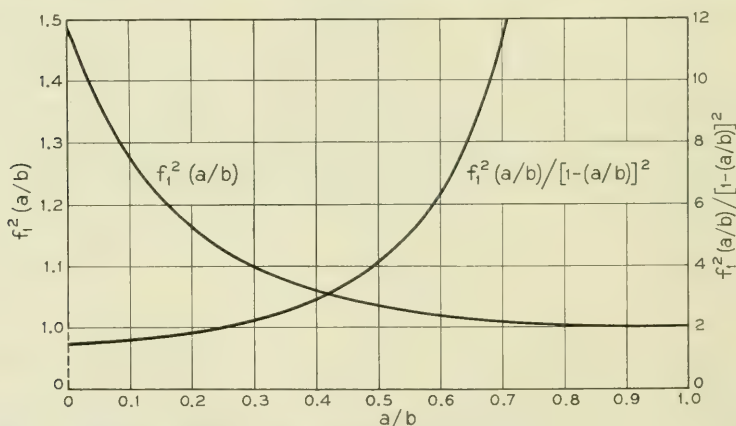


Fig. 12—Curves related to the function

$$f_1^2(a/b) = (b - a)^2 \chi_1^2 / \pi^2,$$

where

$$J_1(\chi_1 a) N_1(\chi_1 b) - J_1(\chi_1 b) N_1(\chi_1 a) = 0.$$

Usually we shall be able to apply these formulas down to frequencies of a few $\text{kc} \cdot \text{sec}^{-1}$.

The field components of the principal mode in a plane Clogston 2 with infinitesimally thin laminae and high-impedance boundaries at $y = \pm \frac{1}{2}a$ are given by equations (271), on substituting for Γ_l from (287). We have, approximately,

$$\begin{aligned} H_x &= H_0 \cos \frac{\pi y}{a} e^{-\gamma z}, \\ \bar{E}_y &= -\sqrt{\frac{\bar{\mu}}{\bar{\epsilon}}} H_0 \cos \frac{\pi y}{a} e^{-\gamma z}, \\ E_z &= \frac{\pi}{\bar{g}a} H_0 \sin \frac{\pi y}{a} e^{-\gamma z}, \end{aligned} \quad (297)$$

where H_0 is an arbitrary amplitude factor, and in the coefficient of the expression for \bar{E}_y we have replaced γ by its approximate value $i\omega\sqrt{\bar{\mu}\bar{\epsilon}}$. The bars have been omitted from H_x and E_z since these field components are continuous at the boundaries of the laminae.

The potential difference between any two points in the same transverse plane is the integral of $-\bar{E}_y$ between the points. In particular, the total potential difference between the upper and lower sheaths is

$$V = -\int_{-\frac{1}{2}a}^{\frac{1}{2}a} \bar{E}_y dy = \frac{2a}{\pi} \sqrt{\frac{\bar{\mu}}{\bar{\epsilon}}} H_0 e^{-\gamma z}. \quad (298)$$

The average value of the conduction current density \bar{J}_z is

$$\bar{J}_z = \bar{g}E_z = \frac{\pi}{a} H_0 \sin \frac{\pi y}{a} e^{-\gamma z}, \quad (299)$$

and the current per unit width flowing in the positive z -direction in the upper half of the stack is

$$I = \int_0^{\frac{1}{2}a} \bar{J}_z dy = H_0 e^{-\gamma z}, \quad (300)$$

so that the ratio of voltage between the sheaths to total one-way current per unit width is

$$\frac{V}{I} = \frac{2a}{\pi} \sqrt{\frac{\bar{\mu}}{\bar{\epsilon}}}. \quad (301)$$

The fields of the principal mode in a coaxial Clogston 2 with infinitesimally thin laminae and high-impedance boundaries are given by equa-

tions (277), which simplify somewhat if we write $i\chi_1$ for Γ_1 , replace the modified Bessel functions with ordinary Bessel functions according to (292), and remember that H_ϕ must vanish at the high-impedance boundaries. We then get, approximately,

$$\begin{aligned} H_\phi &= H_0 [J_1(\chi_1 b) J_1(\chi_1 \rho) - J_1(\chi_1 b) N_1(\chi_1 \rho)] e^{-\gamma z}, \\ \bar{E}_\rho &= \sqrt{\frac{\bar{\mu}}{\bar{\epsilon}}} H_0 [N_1(\chi_1 b) J_1(\chi_1 \rho) - J_1(\chi_1 b) N_1(\chi_1 \rho)] e^{-\gamma z}, \\ E_z &= \frac{\chi_1}{\bar{g}} H_0 [N_1(\chi_1 b) J_0(\chi_1 \rho) - J_1(\chi_1 b) N_0(\chi_1 \rho)] e^{-\gamma z}, \end{aligned} \quad (302)$$

where H_0 is an arbitrary amplitude factor.

The potential difference between any two points in the same transverse plane is the integral of $-\bar{E}_\rho$ between the points. Thus the total potential difference between the core and the outer sheath is

$$\begin{aligned} V &= - \int_a^b \bar{E}_\rho d\rho \\ &= \sqrt{\frac{\bar{\mu}}{\bar{\epsilon}}} \frac{H_0}{\chi_1} [N_1(\chi_1 b) J_0(\chi_1 \rho) - J_1(\chi_1 b) N_0(\chi_1 \rho)]_a^b e^{-\gamma z} \\ &= \sqrt{\frac{\bar{\mu}}{\bar{\epsilon}}} \frac{2H_0}{\pi \chi_1} \left[\frac{J_1(\chi_1 b)}{\chi_1 a J_1(\chi_1 a)} - \frac{1}{\chi_1 b} \right] e^{-\gamma z}, \end{aligned} \quad (303)$$

after some transformations using equation (293) and the well-known identity

$$N_0(x) J_1(x) - N_1(x) J_0(x) = 2/\pi x. \quad (304)$$

The average value of the conduction current density \bar{J}_z is

$$\bar{J}_z = \bar{g} E_z = H_0 \chi_1 [N_1(\chi_1 b) J_0(\chi_1 \rho) - J_1(\chi_1 b) N_0(\chi_1 \rho)] e^{-\gamma z}. \quad (305)$$

The current reverses at $\rho = c$, where $a < c < b$ and c satisfies

$$N_1(\chi_1 b) J_0(\chi_1 c) - J_1(\chi_1 b) N_0(\chi_1 c) = 0; \quad (306)$$

hence the value of c may be found with the aid of a table of Bessel functions or from plotted curves.¹⁹ The total one-way current in the outer part of the stack is

$$\begin{aligned} I &= 2\pi \int_c^b \bar{J}_z \rho d\rho \\ &= 2\pi c H_0 [J_1(\chi_1 b) N_1(\chi_1 c) - N_1(\chi_1 b) J_1(\chi_1 c)] e^{-\gamma z} \\ &= - \frac{4H_0 J_1(\chi_1 b)}{\chi_1 J_0(\chi_1 c)} e^{-\gamma z}, \end{aligned} \quad (307)$$

¹⁹ Reference 18, p. 208.

where in the last step we have made use of (304) and (306). The ratio of voltage across the stack to total one-way current is, from (303) and (307),

$$\frac{V}{I} = \sqrt{\frac{\bar{\mu}}{\bar{\epsilon}}} \frac{J_0(\chi_1 c)}{2\pi} \left[\frac{1}{\chi_1 b J_1(\chi_1 b)} - \frac{1}{\chi_1 a J_1(\chi_1 a)} \right]. \quad (308)$$

If there is no inner core, so that $a = 0$, the expressions which we have just derived become indeterminate forms, and it is simplest to make an independent calculation of the fields for this special case. The Neumann functions are now excluded because of their singularity at $\rho = 0$, and the condition (293) is replaced by

$$J_1(\chi b) = 0, \quad (309)$$

from which

$$\chi_1 = 3.8317/b. \quad (310)$$

The expressions for the fields are

$$\begin{aligned} H_\phi &= H_0 J_1(\chi_1 \rho) e^{-\gamma z}, \\ \bar{E}_\rho &= \sqrt{\frac{\bar{\mu}}{\bar{\epsilon}}} H_0 J_1(\chi_1 \rho) e^{-\gamma z}, \\ E_z &= \frac{\chi_1}{\bar{g}} H_0 J_0(\chi_1 \rho) e^{-\gamma z}, \end{aligned} \quad (311)$$

where H_0 is now a different arbitrary amplitude factor.

The total potential difference across the stack becomes, after putting in numerical values,

$$V = - \int_0^a \bar{E}_\rho d\rho = 0.3661 \sqrt{\bar{\mu}/\bar{\epsilon}} H_0 b e^{-\gamma z}. \quad (312)$$

The conduction current density is

$$\bar{J}_z = \bar{g} E_z = \chi_1 H_0 J_0(\chi_1 \rho) e^{-\gamma z}; \quad (313)$$

and \bar{J}_z changes sign at

$$J_0(\chi_1 c) = 0, \quad c = 2.4048/\chi_1 = 0.6276b. \quad (314)$$

The total one-way current is

$$I = 2\pi c H_0 J_1(\chi_1 c) e^{-\gamma z} = 2.047 H_0 b e^{-\gamma z}, \quad (315)$$

and the ratio of total voltage to total current is

$$V/I = 0.1788 \sqrt{\bar{\mu}/\bar{\epsilon}}. \quad (316)$$

The fields of the principal mode in both plane and coaxial Clogston 2 lines will be plotted in the next section, when we shall also be able to show the fields in various transition structures between the extreme Clogston 1 and the complete Clogston 2.

As a numerical example, let us compare the attenuation constant of a conventional coaxial cable with that of a completely filled Clogston 2 cable of the same size. If a and b denote the radii of the inner and outer conductors of a conventional coaxial cable of optimum proportions ($b/a = 3.5911$), then at frequencies high enough to give a well-developed skin effect on both conductors, the attenuation constant is given by equation (151) of Section IV, namely

$$\alpha = \frac{1.796}{\eta_0 g_1 \delta_1 b}, \quad (317)$$

where η_0 is the intrinsic impedance of the main dielectric, which may be air. On the other hand, the attenuation constant of a Clogston 2 cable of outer radius b , with infinitesimally thin laminae and no inner core, is, from equations (282), (291), and (310),

$$\alpha = \frac{7.341}{\sqrt{\bar{\mu}/\bar{\epsilon}} \bar{g} b^2}. \quad (318)$$

It will be shown in the next section that for infinitesimally thin laminae whose permeabilities are all equal, the optimum value of θ is $2/3$. Assuming no magnetic materials and setting $\theta = 2/3$, we find that the ratio of the attenuation constant α_c of the Clogston cable to the attenuation constant α_s of an *air-filled* standard coaxial cable of the same size, made of the same conducting material, is

$$\alpha_c/\alpha_s = 10.62 \sqrt{\epsilon_{2r}} \delta_1/b. \quad (319)$$

For copper conductors, δ_1 is given by equation (78) of Section III, and the crossover frequency above which the Clogston cable has a lower attenuation constant than the standard coaxial cable turns out to be

$$f_{M0} = 763.5 \epsilon_{2r}/b_{\text{mils}}^2, \quad (320)$$

where frequency is measured in $\text{Mc} \cdot \text{sec}^{-1}$ and the radius of the cable in mils. We also note that at the crossover frequency the electrical radius of the inner conductor of the standard coaxial is $2.96 \sqrt{\epsilon_{2r}} \delta_1$, so that the use of equation (317) for α_s appears to be (barely) justified. Applying equation (320) to an ideal Clogston 2 cable of outer diameter 0.375 inches, excluding the sheath, with copper conductors, polyethylene

insulation, and no inner core, we have

$$b = 187.5 \text{ mils}, \quad \epsilon_{2r} = 2.26, \quad (321)$$

and the crossover frequency is about $50 \text{ kc} \cdot \text{sec}^{-1}$.

It must be emphasized that several factors which have not yet been taken into account will conspire to reduce the practical improvement in transmission that can be obtained with a Clogston 2 cable. As we have already seen for Clogston 1 lines in Part I, the effect of finite lamina thickness in a Clogston 2 will be to cause the attenuation constant to increase with increasing frequency, and ultimately to become higher than the attenuation constant of a conventional coaxial cable. Dissipation in the insulating layers may also contribute appreciably to the total loss at the upper end of the frequency band. Perhaps most important of all, the average electrical properties of the laminated medium must be held extremely uniform across the stack, or the field pattern of the principal mode will be distorted and its attenuation constant correspondingly increased. In later sections we shall discuss these effects, in order to estimate the stringency of the requirements on a physical Clogston cable if its factor of improvement over a conventional cable is to approximate closely to the theoretical limit given, for example, by equation (319).

IX. PARTIALLY FILLED CLOGSTON LINES. OPTIMUM PROPORTIONS FOR PRINCIPAL MODE

The distinction which has heretofore been made between Clogston 1 and Clogston 2 lines is rather artificial, inasmuch as both structures are limiting cases of the general Clogston-type line in which an arbitrary fraction of the total space is occupied by laminated material and the rest by an isotropic main dielectric. We shall now consider the modes which can propagate in a general partially filled line, restricting ourselves for simplicity to stacks of infinitesimally thin layers backed by high-impedance walls. Under these assumptions we first set up equations which must be satisfied by the propagation constants and the fields of all modes having only H_x , E_y , E_z or H_ϕ , E_ρ , E_z field components in a partially filled Clogston line, and then proceed to a study of the lowest or principal mode. We exhibit field plots for this mode at various stages of the transition between the extreme Clogston 1 and the complete Clogston 2 geometry, and investigate the conditions under which the attenuation constant passes through a minimum as the space occupied by the stacks is increased. This leads to the determination of certain optimum proportions for a line intended to transmit the principal mode.

In Section X we shall give a similar but briefer treatment of the various higher modes which can exist in partially or completely filled Clogston lines.

The notation for the parallel-plane line is established in Fig. 5 of Part I. The stacks are bounded externally by high-impedance sheaths at $y = \pm \frac{1}{2}a$, while the main dielectric is bounded by the planes $y = \pm \frac{1}{2}b = \pm (\frac{1}{2}a - s)$. No restrictions are placed on the relative thicknesses b and s of the main dielectric and the stacks. The average electrical constants of the stacks are $\bar{\epsilon}$, $\bar{\mu}$, and \bar{g} , while the electrical constants ϵ_0 and μ_0 of the main dielectric are assumed to satisfy Clogston's condition (102) but are otherwise arbitrary.

As in Section II, the modes may be divided into two classes, according to whether H_x is an even function or an odd function about the center plane $y = 0$. The normal surface impedance $Z(\gamma)$ looking into either stack may be obtained from equation (92) of Section III; if the impedance of the outer sheath is effectively infinite we have

$$Z(\gamma) = K \coth \Gamma_t s = (\Gamma_t / \bar{g}) \coth \Gamma_t s, \quad (322)$$

where

$$\Gamma_t = \left[\frac{i\bar{g}}{\omega\bar{\epsilon}} (\omega^2 \bar{\mu}\bar{\epsilon} + \gamma^2) \right]^{\frac{1}{2}}. \quad (323)$$

Substituting for $Z(\gamma)$ into equations (11) and (13) of Section II, we find that the impedance-matching conditions become

$$\tanh \frac{1}{2}\kappa_0 b \tanh \Gamma_t s = - \frac{i\omega\epsilon_0}{\bar{g}} \frac{\Gamma_t}{\kappa_0}, \quad (324)$$

$$\coth \frac{1}{2}\kappa_0 b \tanh \Gamma_t s = - \frac{i\omega\epsilon_0}{\bar{g}} \frac{\Gamma_t}{\kappa_0}, \quad (325)$$

for the even and odd modes respectively, where

$$\kappa_0 = (\sigma_0^2 - \gamma^2)^{\frac{1}{2}} = (-\omega^2 \mu_0 \epsilon_0 - \gamma^2)^{\frac{1}{2}}. \quad (326)$$

From (323) we have

$$\gamma^2 = -\omega^2 \bar{\mu}\bar{\epsilon} - (i\omega\bar{\epsilon}/\bar{g})\Gamma_t^2, \quad (327)$$

and so from (326),

$$\kappa_0^2 = -\omega^2 (\mu_0 \epsilon_0 - \bar{\mu}\bar{\epsilon}) + (i\omega\bar{\epsilon}/\bar{g})\Gamma_t^2. \quad (328)$$

If Clogston's condition is satisfied, namely

$$\mu_0 \epsilon_0 = \bar{\mu}\bar{\epsilon}, \quad (329)$$

then

$$\kappa_0^2 = (i\omega\bar{\epsilon} \bar{g}) \Gamma_\ell^2, \quad \kappa_0 = \sqrt{i\omega\bar{\epsilon}/\bar{g}} \Gamma_\ell, \quad (330)$$

and the equations for the even and odd modes become, respectively,

$$\tanh \frac{1}{2} \sqrt{i\omega\bar{\epsilon}/\bar{g}} \Gamma_\ell b \tanh \Gamma_\ell s = -\frac{\epsilon_0}{\bar{\epsilon}} \sqrt{\frac{i\omega\bar{\epsilon}}{\bar{g}}} = -\frac{\bar{\mu}}{\mu_0} \sqrt{\frac{i\omega\bar{\epsilon}}{\bar{g}}}, \quad (331)$$

$$\coth \frac{1}{2} \sqrt{i\omega\bar{\epsilon}/\bar{g}} \Gamma_\ell b \tanh \Gamma_\ell s = -\frac{\epsilon_0}{\bar{\epsilon}} \sqrt{\frac{i\omega\bar{\epsilon}}{\bar{g}}} = -\frac{\bar{\mu}}{\mu_0} \sqrt{\frac{i\omega\bar{\epsilon}}{\bar{g}}}. \quad (332)$$

For reference we shall now write down the field components of the various modes. The fields in the main dielectric are given by equations (8) and (12) of Section II, while the fields in the stacks may be obtained without difficulty if we recall that the tangential field components must be continuous at the inner boundary of each stack and that the tangential magnetic field must vanish at the high-impedance surface which forms the outer boundary of the stack.

Taking the even modes first, we have for the fields in the main dielectric,

$$\begin{aligned} H_x &= H_0 \operatorname{ch} \kappa_0 y e^{-\gamma z}, \\ E_y &= -H_0 \frac{\gamma}{i\omega\epsilon_0} \operatorname{ch} \kappa_0 y e^{-\gamma z}, \\ E_z &= -H_0 \frac{\kappa_0}{i\omega\epsilon_0} \operatorname{sh} \kappa_0 y e^{-\gamma z}, \end{aligned} \quad (333)$$

for $-\frac{1}{2}b \leq y \leq \frac{1}{2}b$, where H_0 is an arbitrary amplitude factor, γ and κ_0 are given in terms of Γ_ℓ by (327) and (330), and Γ_ℓ satisfies (331). The fields in the stacks are

$$\begin{aligned} H_x &= H_0 \frac{\operatorname{ch} \frac{1}{2}\kappa_0 b}{\operatorname{sh} \Gamma_\ell s} \operatorname{sh} \Gamma_\ell (\tfrac{1}{2}a \mp y) e^{-\gamma z}, \\ \bar{E}_y &= -H_0 \frac{\gamma}{i\omega\bar{\epsilon}} \frac{\operatorname{ch} \frac{1}{2}\kappa_0 b}{\operatorname{sh} \Gamma_\ell s} \operatorname{sh} \Gamma_\ell (\tfrac{1}{2}a \mp y) e^{-\gamma z}, \\ E_z &= \pm H_0 \frac{\Gamma_\ell}{\bar{g}} \frac{\operatorname{ch} \frac{1}{2}\kappa_0 b}{\operatorname{sh} \Gamma_\ell s} \operatorname{ch} \Gamma_\ell (\tfrac{1}{2}a \mp y) e^{-\gamma z}, \end{aligned} \quad (334)$$

for $\frac{1}{2}b \leq |y| \leq \frac{1}{2}a$, where in case of ambiguous signs the upper sign is to be associated with the upper stack ($y > 0$) and the lower sign with the lower stack ($y < 0$). The continuity of E_z at $y = \pm \frac{1}{2}b$ is a consequence of equation (324) or (331).

For the odd modes, the fields in the main dielectric are

$$\begin{aligned} H_x &= H_0 \operatorname{sh} \kappa_0 y e^{-\gamma z}, \\ E_y &= -H_0 \frac{\gamma}{i\omega\epsilon_0} \operatorname{sh} \kappa_0 y e^{-\gamma z}, \\ E_z &= -H_0 \frac{\kappa_0}{i\omega\epsilon_0} \operatorname{ch} \kappa_0 y e^{-\gamma z}, \end{aligned} \quad (335)$$

for $-\frac{1}{2}b \leq y \leq \frac{1}{2}b$, where H_0 is again an arbitrary amplitude factor and γ and κ_0 are defined as before in terms of Γ_ℓ , which is now a root of (332). The fields in the stacks are

$$\begin{aligned} H_x &= \pm H_0 \frac{\operatorname{sh} \frac{1}{2}\kappa_0 b}{\operatorname{sh} \Gamma_\ell s} \operatorname{sh} \Gamma_\ell (\tfrac{1}{2}a \mp y) e^{-\gamma z}, \\ \bar{E}_y &= \mp H_0 \frac{\gamma}{i\omega\epsilon} \frac{\operatorname{sh} \frac{1}{2}\kappa_0 b}{\operatorname{sh} \Gamma_\ell s} \operatorname{sh} \Gamma_\ell (\tfrac{1}{2}a \mp y) e^{-\gamma z}, \\ E_z &= +H_0 \frac{\Gamma_\ell}{\bar{g}} \frac{\operatorname{sh} \frac{1}{2}\kappa_0 b}{\operatorname{sh} \Gamma_\ell s} \operatorname{ch} \Gamma_\ell (\tfrac{1}{2}a \mp y) e^{-\gamma z}, \end{aligned} \quad (336)$$

for $\frac{1}{2}b \leq |y| \leq \frac{1}{2}a$, where again the upper signs refer to the upper stack and the lower signs to the lower stack. The continuity of E_z at $y = \pm \frac{1}{2}b$ is now a consequence of equation (325) or (332).

The notation for the partially filled coaxial cable is shown in Fig. 6 of Part I, where as before we assume that the laminae are infinitesimally thin, the boundary impedances are effectively infinite, and the main dielectric satisfies Clogston's condition. The radius of the inner core is a and that of the outer sheath is b , while the stack thicknesses are s_1 and s_2 respectively; but no restrictions, other than obvious geometrical limitations, are placed on the relative values of a , b , s_1 , and s_2 . The inner and outer radii of the main dielectric are denoted by $\rho_1 (= a + s_1)$ and $\rho_2 (= b - s_2)$ respectively.

The boundary conditions at the surfaces of the main dielectric will be satisfied, as in Section II, by matching radial impedances at the stack-dielectric interfaces. If the impedance Z_a looking into the core at $\rho = a$ is effectively infinite, then the impedance looking into the inner stack at ρ_1 is given by equation (98) of Section III to be

$$Z_1 = \frac{\Gamma_\ell K_0(\Gamma_\ell \rho_1) I_1(\Gamma_\ell a) + K_1(\Gamma_\ell a) I_0(\Gamma_\ell \rho_1)}{\bar{g} K_1(\Gamma_\ell a) I_1(\Gamma_\ell \rho_1) - K_1(\Gamma_\ell \rho_1) I_1(\Gamma_\ell a)}. \quad (337)$$

Similarly, if the sheath impedance Z_b is infinite, then looking into the

outer stack at ρ_2 we have

$$Z_2 = \frac{\Gamma_\ell K_0(\Gamma_\ell \rho_2) I_1(\Gamma_\ell b) + K_1(\Gamma_\ell b) I_0(\Gamma_\ell \rho_2)}{\bar{g} K_1(\Gamma_\ell \rho_2) I_1(\Gamma_\ell b) - K_1(\Gamma_\ell b) I_1(\Gamma_\ell \rho_2)}. \quad (338)$$

The condition that the radial impedances shall be matched at the surfaces of the main dielectric is given by equation (38) of Section II, which takes the form

$$\frac{\kappa_0 K_0(\kappa_0 \rho_1) + i\omega \epsilon_0 Z_1 K_1(\kappa_0 \rho_1)}{\kappa_0 I_0(\kappa_0 \rho_1) - i\omega \epsilon_0 Z_1 I_1(\kappa_0 \rho_1)} = \frac{\kappa_0 K_0(\kappa_0 \rho_2) - i\omega \epsilon_0 Z_2 K_1(\kappa_0 \rho_2)}{\kappa_0 I_0(\kappa_0 \rho_2) + i\omega \epsilon_0 Z_2 I_1(\kappa_0 \rho_2)}, \quad (339)$$

where κ_0 is related to Γ_ℓ by equation (330). If we substitute the expressions (337) and (338) for Z_1 and Z_2 into (339), we have a single equation whose roots in Γ_ℓ correspond to all the circular transverse magnetic modes on the coaxial Clogston line. The propagation constant γ of each mode is given in terms of Γ_ℓ by equation (327).

Once the boundary conditions have been satisfied for a particular mode by a suitable determination of Γ_ℓ , it is a routine matter to obtain the field components for this mode. In the main dielectric the fields are of the form given by equations (33) of Section II. Hence for $\rho_1 \leq \rho \leq \rho_2$ we have

$$\begin{aligned} H_\phi &= [AI_1(\kappa_0 \rho) + BK_1(\kappa_0 \rho)]e^{-\gamma z}, \\ E_\rho &= \frac{\gamma}{i\omega \epsilon_0} [AI_1(\kappa_0 \rho) + BK_1(\kappa_0 \rho)]e^{-\gamma z}, \\ E_z &= \frac{\kappa_0}{i\omega \epsilon_0} [AI_0(\kappa_0 \rho) - BK_0(\kappa_0 \rho)]e^{-\gamma z}, \end{aligned} \quad (340)$$

where one of the constants A and B is arbitrary, but the ratio A/B must be taken equal to either side of equation (339). The fields in the stacks are of the form of equations (277) of Section VIII, where the constants are to be determined so that $H_\phi = 0$ at $\rho = a$ and $\rho = b$, and so that the tangential field components are continuous at ρ_1 and ρ_2 . Imposing these conditions, we find that in the inner stack, for $a \leq \rho \leq \rho_1$,

$$\begin{aligned} H_\phi &= C[K_1(\Gamma_\ell a)I_1(\Gamma_\ell \rho) - I_1(\Gamma_\ell a)K_1(\Gamma_\ell \rho)]e^{-\gamma z}, \\ \bar{E}_\rho &= \frac{\gamma}{i\omega \bar{\epsilon}} C[K_1(\Gamma_\ell a)I_1(\Gamma_\ell \rho) - I_1(\Gamma_\ell a)K_1(\Gamma_\ell \rho)]e^{-\gamma z}, \\ E_z &= \frac{\Gamma_\ell}{\bar{g}} C[K_1(\Gamma_\ell a)I_0(\Gamma_\ell \rho) + I_1(\Gamma_\ell a)K_0(\Gamma_\ell \rho)]e^{-\gamma z}, \end{aligned} \quad (341)$$

where

$$C = \frac{AI_1(\kappa_0\rho_1) + BK_1(\kappa_0\rho_1)}{K_1(\Gamma_\ell a)I_1(\Gamma_\ell\rho_1) - I_1(\Gamma_\ell a)K_1(\Gamma_\ell\rho_1)}. \quad (342)$$

In the outer stack, for $\rho_2 \leq \rho \leq b$,

$$\begin{aligned} H_\phi &= D[K_1(\Gamma_\ell b)I_1(\Gamma_\ell\rho) - I_1(\Gamma_\ell b)K_1(\Gamma_\ell\rho)]e^{-\gamma z}, \\ \bar{E}_\rho &= \frac{\gamma}{i\omega\bar{\epsilon}} D[K_1(\Gamma_\ell b)I_1(\Gamma_\ell\rho) - I_1(\Gamma_\ell b)K_1(\Gamma_\ell\rho)]e^{-\gamma z}, \\ E_z &= \frac{\Gamma_\ell}{\bar{g}} D[K_1(\Gamma_\ell b)I_0(\Gamma_\ell\rho) + I_1(\Gamma_\ell b)K_0(\Gamma_\ell\rho)]e^{-\gamma z}, \end{aligned} \quad (343)$$

where

$$D = \frac{AI_1(\kappa_0\rho_2) + BK_1(\kappa_0\rho_2)}{K_1(\Gamma_\ell b)I_1(\Gamma_\ell\rho_2) - I_1(\Gamma_\ell b)K_1(\Gamma_\ell\rho_2)}. \quad (344)$$

For the remainder of the present section we shall confine our attention to the principal mode. In the parallel-plane line this mode corresponds to the lowest root in Γ_ℓ of equation (331), that is,

$$\tanh \frac{1}{2}\sqrt{i\omega\bar{\epsilon}/\bar{g}} \Gamma_\ell b \tanh \Gamma_\ell s = -\frac{\bar{\mu}}{\mu_0} \sqrt{i\omega\bar{\epsilon}/\bar{g}}. \quad (345)$$

We note that the right side of the equation is very small compared to unity, being of the order of the square root of the ratio of displacement current density in the insulators to conduction current density in the conductors, and also that the coefficient of Γ_ℓ in the first factor on the left will under all ordinary conditions be much smaller than the coefficient of Γ_ℓ in the second factor. Hence in seeking the lowest root we are justified in replacing the first hyperbolic tangent on the left side of (345) by its argument, so that the equation becomes

$$\Gamma_\ell s \tanh \Gamma_\ell s = -\frac{\bar{\mu}}{\mu_0} \frac{2s}{b}. \quad (346)$$

If we now let

$$\Gamma_\ell^2 = -\chi^2, \quad \Gamma_\ell = i\chi, \quad (347)$$

we obtain

$$\chi s \tan \chi s = \frac{\bar{\mu}}{\mu_0} \frac{2s}{b}. \quad (348)$$

Since the right side of (348) is a positive real constant, the equation has

exactly one root in the interval $0 < \chi s \leq \frac{1}{2}\pi$, which may most easily be found from a table²⁰ of the function $x \tan x$. If we call this root χ_1 , equation (327) for the propagation constant γ becomes

$$\gamma^2 = -\omega^2 \bar{\mu} \bar{\epsilon} + (i\omega \bar{\epsilon}' \bar{g}) \chi_1^2; \quad (349)$$

and on taking the square root by the binomial theorem we find for the attenuation and phase constants of the principal mode,

$$\alpha = \frac{\chi_1^2}{2\sqrt{\bar{\mu}/\bar{\epsilon}} \bar{g}}, \quad (350)$$

$$\beta = \omega \sqrt{\bar{\mu} \bar{\epsilon}}. \quad (351)$$

It is easy to verify that (350) reduces to the expressions previously obtained for the attenuation constants of Clogston 1 and Clogston 2 lines in the limiting cases $s \ll \frac{1}{2}a$ and $s = \frac{1}{2}a$ respectively. If $s \ll \frac{1}{2}a$, (348) gives

$$\chi_1^2 = \frac{2(\bar{\mu}/\mu_0)}{bs}, \quad (352)$$

so that from (350), on making use of Clogston's condition,

$$\alpha = \frac{(\bar{\mu}/\mu_0)}{\sqrt{\bar{\mu}/\bar{\epsilon}} \bar{g}bs} = \frac{1}{\sqrt{\mu_0/\epsilon_0} \bar{g}bs}, \quad (353)$$

which agrees with equation (110) of Section IV. If $s = \frac{1}{2}a$, so that $b = 0$, then from (348),

$$\chi_1 = \frac{1}{2}\pi/s = \pi/a, \quad (354)$$

and (350) becomes

$$\alpha = \frac{\pi^2}{2\sqrt{\bar{\mu}/\bar{\epsilon}} \bar{g}a^2}, \quad (355)$$

which is the same as equation (288) of the preceding section.

The general expressions (333) and (334) for the fields in a plane Clogston line with infinitesimally thin laminae and high-impedance walls simplify considerably for the principal mode, since κ_0 is so small that to a good approximation for $|y| \leq \frac{1}{2}b$ we may replace $\text{sh } \kappa_0 y$ by $\kappa_0 y$ and $\text{ch } \kappa_0 y$ by unity. We then have, in the main dielectric,

²⁰ See for example Reference 18, Addenda, pp. 32-35.

$$\begin{aligned}
 H_x &\approx H_0 e^{-\gamma z}, \\
 E_y &\approx -\sqrt{\frac{\mu_0}{\epsilon_0}} H_0 e^{-\gamma z}, \\
 E_z &\approx \frac{\mu_0 \chi_1^2}{\bar{\mu} \bar{g}} H_0 y e^{-\gamma z},
 \end{aligned} \tag{356}$$

for $-\frac{1}{2}b \leq y \leq \frac{1}{2}b$, where H_0 is an arbitrary amplitude factor. The fields in the stacks are

$$\begin{aligned}
 H_x &\approx H_0 \frac{\sin \chi_1(\frac{1}{2}a \mp y)}{\sin \chi_1 s} e^{-\gamma z}, \\
 \bar{E}_y &\approx -\sqrt{\frac{\bar{\mu}}{\bar{\epsilon}}} H_0 \frac{\sin \chi_1(\frac{1}{2}a \mp y)}{\sin \chi_1 s} e^{-\gamma z}, \\
 E_z &\approx \pm \frac{\chi_1}{\bar{g}} H_0 \frac{\cos \chi_1(\frac{1}{2}a \mp y)}{\sin \chi_1 s} e^{-\gamma z},
 \end{aligned} \tag{357}$$

for $\frac{1}{2}b \leq |y| \leq \frac{1}{2}a$, where the upper signs refer to the upper stack and the lower signs to the lower stack. The potential and current distribu-

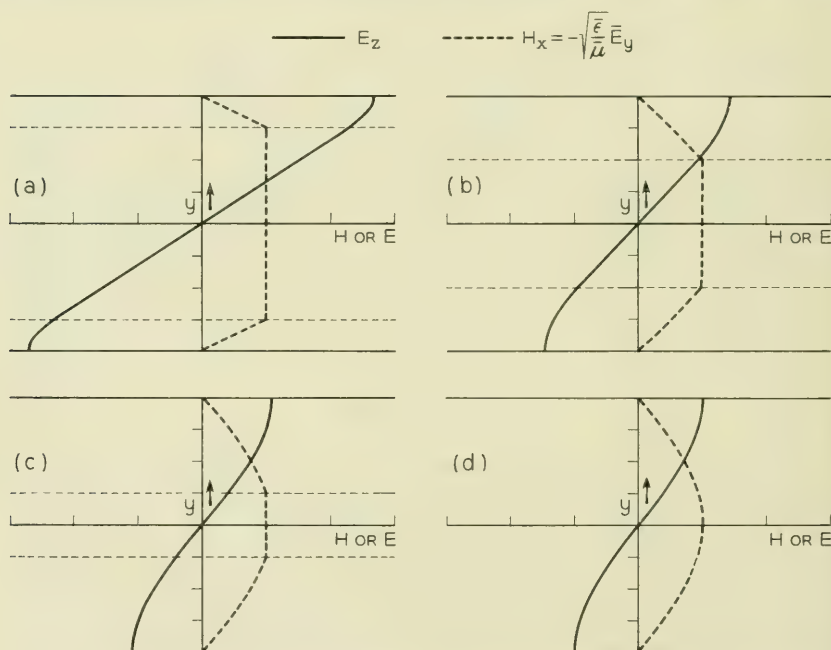


Fig. 13—Fields of principal mode in partially and completely filled plane Clogston lines with $\mu_0 = \bar{\mu}$, $\epsilon_0 = \bar{\epsilon}$.

tions may easily be obtained, if desired, from the expressions for the field components.

As a numerical example we have plotted in Fig. 13 the fields of the principal mode for plane transmission lines in which the stacks fill respectively one-quarter, one-half, three-quarters, and all of the total available space. For simplicity we have taken $\mu_0 = \bar{\mu}$, $\epsilon_0 = \bar{\epsilon}$, and normalized the fields so that the total one-way current is the same in all four cases. The average current density is of course $\bar{g}E_z$ in the stacks and zero in the main dielectric. The first case approximates most nearly to the extreme Clogston 1 line discussed in Part I, while the last case is the complete Clogston 2, and the intermediate cases show the transition between these two structures. The following table gives $\chi_1 s$ as a function of the fraction $s/\frac{1}{2}a$ of the total space filled by the stacks, and also the quantity $(\chi_1 a/\pi)^2$, which is equal to the ratio of the attenuation constant of the line to the attenuation constant of the complete Clogston 2.

$s/\frac{1}{2}a$	$\chi_1 s \tan \chi_1 s$	$\chi_1 s$	$(\chi_1 a/\pi)^2$
$\frac{1}{4}$	$\frac{1}{3}$	0.5471	1.941
$\frac{1}{2}$	1	0.8603	1.200
$\frac{3}{4}$	3	1.1924	1.024
1	∞	1.5708	1.000

The principal mode in a coaxial Clogston cable corresponds to the lowest root in Γ_t of equation (339). For the principal mode we are justified in assuming that in the main dielectric

$$|\kappa_0 \rho| \ll 1, \quad (358)$$

so that the Bessel functions occurring in equation (339) may be replaced by their approximate values for small argument. We thus find that, to a very good approximation, equation (339) reduces to the same form as equation (41) of Section II, namely

$$\frac{\kappa_0^2}{i\omega\epsilon_0} \log \frac{\rho_2}{\rho_1} = -\left(\frac{Z_1}{\rho_1} + \frac{Z_2}{\rho_2}\right), \quad (359)$$

where the stack impedances Z_1 and Z_2 are given by (337) and (338). If as before we let

$$\Gamma_t^2 = -\chi^2, \quad \Gamma_t = i\chi, \quad (360)$$

then from (330) κ_0^2 becomes

$$\kappa_0^2 = -(i\omega\bar{\epsilon}/\bar{g})\chi^2; \quad (361)$$

and replacing the modified Bessel functions in (337) and (338) with ordinary Bessel functions according to equations (292) of Section VIII, we obtain

$$Z_1 = \frac{\chi}{\bar{g}} \frac{J_1(\chi a)N_0(\chi \rho_1) - N_1(\chi a)J_0(\chi \rho_1)}{J_1(\chi a)N_1(\chi \rho_1) - N_1(\chi a)J_1(\chi \rho_1)}, \quad (362)$$

$$Z_2 = \frac{\chi}{\bar{g}} \frac{J_1(\chi b)N_0(\chi \rho_2) - N_1(\chi b)J_0(\chi \rho_2)}{J_1(\chi \rho_2)N_1(\chi b) - N_1(\chi \rho_2)J_1(\chi b)}. \quad (363)$$

Substituting (361), (362), and (363) into (359) and setting $\bar{\epsilon}/\epsilon_0 = \mu_0/\bar{\mu}$, we get after a little rearrangement,

$$\begin{aligned} \frac{1}{\chi \rho_1} \frac{J_1(\chi a)N_0(\chi \rho_1) - N_1(\chi a)J_0(\chi \rho_1)}{J_1(\chi a)N_1(\chi \rho_1) - N_1(\chi a)J_1(\chi \rho_1)} \\ + \frac{1}{\chi \rho_2} \frac{J_1(\chi b)N_0(\chi \rho_2) - N_1(\chi b)J_0(\chi \rho_2)}{J_1(\chi \rho_2)N_1(\chi b) - N_1(\chi \rho_2)J_1(\chi b)} = \frac{\mu_0}{\bar{\mu}} \log \frac{\rho_2}{\rho_1}. \end{aligned} \quad (364)$$

If χ_1 is the smallest positive root of equation (364), then the attenuation and phase constants of the principal mode are

$$\alpha = \frac{\chi_1^2}{2\sqrt{\bar{\mu}/\bar{\epsilon}} \bar{g}}, \quad (365)$$

$$\beta = \omega \sqrt{\bar{\mu}\bar{\epsilon}}. \quad (366)$$

These expressions for α and β are of exactly the same form as equations (295) and (296) for a complete Clogston 2, except that χ_1 is now determined from equation (364) instead of equation (293). It is easy to show that (364) reduces to (293) if there is no main dielectric, that is, if $\rho_1 = \rho_2$.

For any given values of the four ratios a/b , ρ_1/b , ρ_2/b , and $\mu_0/\bar{\mu}$, equation (364) may be solved for $\chi_1 b$ by numerical or graphical methods. However if we wish to examine many cases, so as to investigate the effects of varying some or all of the parameters, a more efficient procedure for finding χ_1 is needed. Such a method is provided by the observation that in spite of the complicated appearance of equation (364), it is really just the equation which determines the eigenvalues in a rather simple two-point boundary-value problem, which is well adapted to solution on a differential analyzer. We digress briefly to formulate this problem.

The differential equations for the fields in the main dielectric can be put in the form of equations (67) of Section III, namely

$$\begin{aligned} d(-\rho H_\phi)/d\rho &= -i\omega\epsilon_0\rho E_z, \\ dE_z/d\rho &= -(\kappa_0^2/i\omega\epsilon_0\rho)(-\rho H_\phi), \end{aligned} \quad (367)$$

where the propagation factor $e^{-\gamma z + i\omega t}$ has been suppressed. On the other hand, equations (270) for the fields in a stack of infinitesimally thin laminae yield

$$\begin{aligned} d(-\rho H_\phi)/d\rho &= -\bar{g}\rho E_z, \\ dE_z/d\rho &= -(\Gamma_t^2/\bar{g}\rho)(-\rho H_\phi), \end{aligned} \quad (368)$$

where Γ_t is given by (323). If we neglect the displacement current in the main dielectric compared with the conduction current in the stacks, replace Γ_t by $i\chi$, express κ_0 in terms of χ by (361), and write $\mu_0/\bar{\mu}$ for $\bar{\epsilon}/\epsilon_0$, we obtain the following equations for the fields in the coaxial Clogston line:

For $a \leq \rho \leq \rho_1$,

$$\begin{aligned} d(-\rho H_\phi)/d\rho &= -\rho(\bar{g}E_z), \\ d(\bar{g}E_z)/d\rho &= (\chi^2/\rho)(-\rho H_\phi); \end{aligned} \quad (369i)$$

while for $\rho_1 \leq \rho \leq \rho_2$,

$$\begin{aligned} d(-\rho H_\phi)/d\rho &= 0, \\ d(\bar{g}E_z)/d\rho &= (\mu_0\chi^2/\bar{\mu}\rho)(-\rho H_\phi); \end{aligned} \quad (369ii)$$

and for $\rho_2 \leq \rho \leq b$,

$$\begin{aligned} d(-\rho H_\phi)/d\rho &= -\rho(\bar{g}E_z), \\ d(\bar{g}E_z)/d\rho &= (\chi^2/\rho)(-\rho H_\phi). \end{aligned} \quad (369iii)$$

The quantities $-\rho H_\phi$ and $\bar{g}E_z$ must be continuous at ρ_1 and ρ_2 ; and the two-point boundary condition at the infinite-impedance surfaces $\rho = a$ and $\rho = b$, namely

$$-aH_\phi(a) = -bH_\phi(b) = 0, \quad (370)$$

determines a sequence of eigenvalues $\chi_1^2, \chi_2^2, \chi_3^2, \dots$, of which the lowest corresponds to the principal mode. It is a routine matter to integrate equations (369) in terms of Bessel functions and logarithms, and to show that the continuity and boundary conditions lead exactly to equation (364).

If equations (369) are set up on a differential analyzer with adjustable values of χ^2 , a/b , ρ_1/b , ρ_2/b , and $\mu_0/\bar{\mu}$, it is a simple procedure to make a few runs with different choices of χ^2 , and so to locate the approximate

value of χ_1^2 which satisfies the boundary conditions for the given values of the other parameters. If additional accuracy is wanted, it is then not too difficult to refine this approximate value by desk computation. The results of quite a number of exploratory calculations which were made on the Laboratories' general purpose analog computer will be shown later in this section.

The fields of the principal mode in the main dielectric of a Clogston cable are given approximately by equations (46) of Section II, namely

$$\begin{aligned} H_\phi &\approx \frac{I}{2\pi\rho} e^{-\gamma z}, \\ E_\rho &\approx \sqrt{\frac{\mu_0}{\epsilon_0}} \frac{I}{2\pi\rho} e^{-\gamma z}, \\ E_z &\approx \frac{I}{2\pi \log(\rho_2/\rho_1)} \left[\frac{Z_1}{\rho_1} \log \frac{\rho_2}{\rho} + \frac{Z_2}{\rho_2} \log \frac{\rho_1}{\rho} \right] e^{-\gamma z}, \end{aligned} \quad (371)$$

for $\rho_1 \leq \rho \leq \rho_2$, where I is an amplitude factor equal to the total current flowing in the positive z -direction in the inner stack, and Z_1 and Z_2 are given by writing χ_1 for χ in (362) and (363) respectively. The fields in the inner stack are

$$\begin{aligned} H_\phi &\approx \frac{I}{2\pi\rho_1} \frac{N_1(\chi_1 a)J_1(\chi_1\rho) - J_1(\chi_1 a)N_1(\chi_1\rho)}{N_1(\chi_1 a)J_1(\chi_1\rho_1) - J_1(\chi_1 a)N_1(\chi_1\rho_1)} e^{-\gamma z}, \\ \bar{E}_\rho &\approx \sqrt{\frac{\bar{\mu}}{\bar{\epsilon}}} \frac{I}{2\pi\rho_1} \frac{N_1(\chi_1 a)J_1(\chi_1\rho) - J_1(\chi_1 a)N_1(\chi_1\rho)}{N_1(\chi_1 a)J_1(\chi_1\rho_1) - J_1(\chi_1 a)N_1(\chi_1\rho_1)} e^{-\gamma z}, \\ E_z &\approx \frac{\chi_1}{\bar{g}} \frac{I}{2\pi\rho_1} \frac{N_1(\chi_1 a)J_0(\chi_1\rho) - J_1(\chi_1 a)N_0(\chi_1\rho)}{N_1(\chi_1 a)J_1(\chi_1\rho_1) - J_1(\chi_1 a)N_1(\chi_1\rho_1)} e^{-\gamma z}, \end{aligned} \quad (372)$$

for $a \leq \rho \leq \rho_1$; while in the outer stack we have

$$\begin{aligned} H_\phi &\approx \frac{I}{2\pi\rho_2} \frac{N_1(\chi_1 b)J_1(\chi_1\rho) - J_1(\chi_1 b)N_1(\chi_1\rho)}{N_1(\chi_1 b)J_1(\chi_1\rho_2) - J_1(\chi_1 b)N_1(\chi_1\rho_2)} e^{-\gamma z}, \\ \bar{E}_\rho &\approx \sqrt{\frac{\bar{\mu}}{\bar{\epsilon}}} \frac{I}{2\pi\rho_2} \frac{N_1(\chi_1 b)J_1(\chi_1\rho) - J_1(\chi_1 b)N_1(\chi_1\rho)}{N_1(\chi_1 b)J_1(\chi_1\rho_2) - J_1(\chi_1 b)N_1(\chi_1\rho_2)} e^{-\gamma z}, \\ E_z &\approx \frac{\chi_1}{\bar{g}} \frac{I}{2\pi\rho_2} \frac{N_1(\chi_1 b)J_0(\chi_1\rho) - J_1(\chi_1 b)N_0(\chi_1\rho)}{N_1(\chi_1 b)J_1(\chi_1\rho_2) - J_1(\chi_1 b)N_1(\chi_1\rho_2)} e^{-\gamma z}, \end{aligned} \quad (373)$$

for $\rho_2 \leq \rho \leq b$. The potential and current distributions may be calculated in the usual way from the fields.

As numerical examples we have plotted in Fig. 14 the fields of the principal mode in two Clogston coaxial cables. Fig. 14(a) shows a

cable in which $\mu_0 = \bar{\mu}$, $\epsilon_0 = \bar{\epsilon}$, and with the dimensions $a = 0.084b$, $\rho_1 = 0.415b$, and $\rho_2 = 0.831b$, these proportions having been found optimum, as discussed below, for a cable without magnetic loading in which the total thickness of both stacks is arbitrarily chosen equal to $\frac{1}{2}b$. Fig. 14(b) shows the fields of a complete Clogston 2 with no inner core, the scale being chosen so that the total one-way current is the same in both cases. The attenuation constant of the first cable is 1.234 times that of the second one. Fig. 14 may be compared with Fig. 13 for the plane geometry, whence it should be possible to visualize approximately the fields of the principal mode in other coaxial structures representing various stages of the transition between the extreme Clogston 1 and the complete Clogston 2 cable.

Now let us consider a Clogston line with infinitesimally thin laminae, having fixed external dimensions and containing only materials with given electrical constants. We may pose two questions: (1) Supposing that for some practical reason the total available thickness of laminated material is also fixed, how should this material be divided between the two stacks to minimize the attenuation constant of the line? (2) Assuming that the total thickness of laminations in the line is at our disposal, what is the optimum amount of laminated material from the standpoint of minimizing the attenuation, and how should this material be distributed in the optimum case?

For plane transmission lines the first question is trivial; the stacks should always be of equal thickness. In a coaxial cable, if the filling ratio $(s_1 + s_2)/b$ is given, the proportions of the cable are completely determined when we specify, for example, the relative radius a/b of the core and the relative thickness $s_1/(s_1 + s_2)$ of the inner stack. The opti-

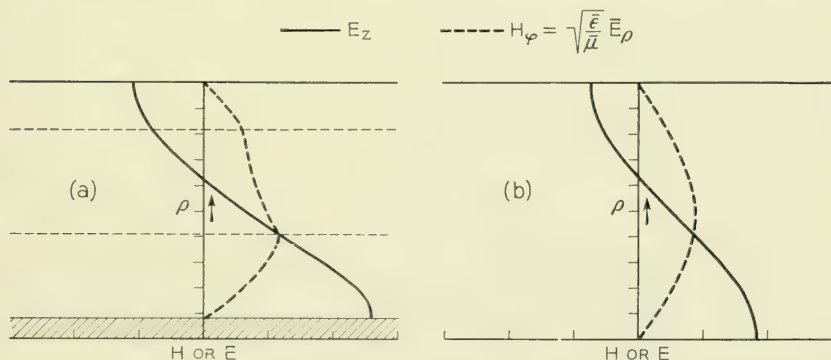


Fig. 14—Fields of principal mode in partially and completely filled coaxial Clogston lines with $\mu_0 = \bar{\mu}$, $\epsilon_0 = \bar{\epsilon}$.

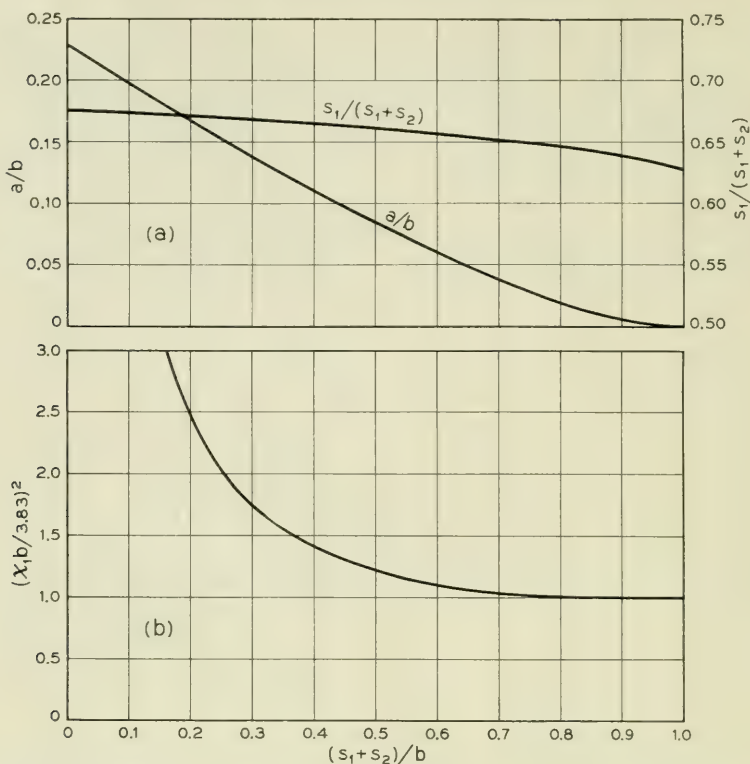


Fig. 15—Relative proportions and relative attenuation constants of optimum Clogston cables with different filling ratios and $\mu_0 = \bar{\mu}$, $\epsilon_0 = \bar{\epsilon}$.

imum values of these two ratios in the extreme Clogston 1 case, where $(s_1 + s_2) \ll b$, have already been given in equations (138) and (139) of Section IV, while in a complete Clogston 2 with $s_1 + s_2 = b$, the stacks should be divided at the radius $0.6276b$, where according to equation (314) the current density is zero. For intermediate filling ratios, with any fixed magnetic loading ratio $\mu_0/\bar{\mu}$, the optimum distribution of laminated material can most easily be found numerically by calculating χ_1^2 or $(\chi_1 b)^2$ for a number of different choices of the ratios a/b and $s_1/(s_1 + s_2)$, and then locating the minimum by double interpolation.

The results of applying this numerical procedure to Clogston cables with various filling ratios and no magnetic loading are plotted in Fig. 15, the necessary values of $\chi_1 b$ having been found on the analog computer and then refined by desk computation. Fig. 15(a) shows the optimum values of a/b and $s_1/(s_1 + s_2)$ as functions of the filling ratio $(s_1 + s_2)/b$, while Fig. 15(b) shows the corresponding value of $(\chi_1 b / 3.83)^2$,

which by equation (365) is proportional to the attenuation constant. We note that the Clogston 2 line with filling ratio unity has the lowest attenuation constant of any cable of the same size without magnetic loading, but that the attenuation constant increases only slowly as the filling ratio decreases, so long as the ratio is greater than about one-half. It also appears that the minimum in $\chi_1 b$, considered as a function of a/b and $s_1/(s_1 + s_2)$ for a fixed filling ratio, is quite broad, which means that in practice one can attain very nearly optimum performance even while deviating somewhat from the optimum proportions.

If the filling ratio is at our disposal, then the solution of the optimum problem is as follows: When there is no magnetic loading of the main dielectric relative to the stacks, that is, when $\mu_0 \leq \bar{\mu}$, then minimum attenuation is obtained with a complete Clogston 2. If on the other hand there is magnetic loading of the main dielectric, so that $\mu_0 > \bar{\mu}$, then minimum attenuation is obtained with a filling ratio less than unity, whose value is a function of the ratio $\mu_0/\bar{\mu}$.

According to equation (350), the attenuation constant of a plane Clogston line is

$$\alpha = \frac{\chi_1^2}{2\sqrt{\bar{\mu}/\epsilon} \bar{g}}, \quad (374)$$

where χ_1 is given by equation (348),

$$\chi_1 \tan \chi_1 s = \frac{2\bar{\mu}}{\mu_0 b} = \frac{\bar{\mu}}{\mu_0(\frac{1}{2}a - s)}. \quad (375)$$

To find the minimum value of χ_1 when a and $\mu_0/\bar{\mu}$ are given, we differentiate (375) with respect to s and set $d\chi_1/ds$ equal to zero. This gives

$$\chi_1^2 \sec^2 \chi_1 s = \frac{\bar{\mu}}{\mu_0(\frac{1}{2}a - s)^2}, \quad (375.5)$$

which when solved simultaneously with equation (375) leads to

$$\sin \chi_1 s = \sqrt{\bar{\mu}/\mu_0}, \quad \chi_1 = \frac{1}{s} \sin^{-1} \sqrt{\bar{\mu}/\mu_0}. \quad (376)$$

Substituting this value of χ_1 into (375) and solving for s in terms of a , we get

$$s = \frac{1}{2}a \frac{\mu_0 \sin^{-1} \sqrt{\bar{\mu}/\mu_0}}{\mu_0 \sin^{-1} \sqrt{\bar{\mu}/\mu_0} + \sqrt{\bar{\mu}(\mu_0 - \bar{\mu})}}, \quad (377)$$

and from (374) the corresponding attenuation constant is

$$\alpha = \frac{2}{\sqrt{\bar{\mu}/\epsilon} \bar{g} a^2} [\sin^{-1} \sqrt{\bar{\mu}/\mu_0} + \sqrt{\bar{\mu}(\mu_0 - \bar{\mu})/\mu_0^2}]^2. \quad (378)$$

As $\mu_0/\bar{\mu}$ increases from unity to very large values, the optimum value of s decreases from $\frac{1}{2}a$ toward $\frac{1}{4}a$, so that the filling ratio decreases from unity toward one-half. If $\mu_0/\bar{\mu} < 1$, equation (376) does not yield a real solution, but the complete Clogston 2 is still the physical structure having the lowest attenuation.

For a coaxial Clogston line without magnetic loading the optimum filling ratio is unity, as we have seen above, while in the presence of magnetic loading a smaller filling ratio is optimum. This filling ratio and the optimum distribution of the laminated material in the cable can be determined by numerical analysis for any given value of $\mu_0/\bar{\mu}$. It is reasonably evident on physical grounds, and can be proved mathematically by a variational argument applied to the lowest eigenvalue of equations (369), that whatever may be the radii ρ_1 and ρ_2 of the main dielectric, the lowest attenuation constant is achieved when $a = 0$, that is, when there is no core inside the inner stack. (This is only a mathematical limit; from a practical standpoint, the use of a small core in the manufacturing process is not likely to make any significant increase in the attenuation of the cable.) For each value of the loading ratio $\mu_0/\bar{\mu}$, therefore, we have merely to minimize the value of $(\chi_1 b)^2$ as a function of the two ratios ρ_1/b and ρ_2/b , which can be done by the double interpolation procedure mentioned earlier. We find that as $\mu_0/\bar{\mu}$ increases from unity to very large values, the optimum value of ρ_1 decreases from $0.6276b$ toward $0.3930b$, while ρ_2 increases from $0.6276b$ toward $0.8226b$, so that the filling ratio decreases from unity toward 0.5704 . The limiting values of ρ_1 and ρ_2 when $\mu_0/\bar{\mu} \gg 1$ are derived from equation (364) by the method shown in Appendix II.

As a numerical example we have considered a Clogston cable with $\mu_0 = 3\bar{\mu}$. The optimum proportions of this cable and the corresponding value of χ_1 are approximately

$$\rho_1 = 0.426b, \quad \rho_2 = 0.796b, \quad \chi_1 = 2.720/b; \quad (379)$$

and the minimum attenuation constant is

$$\alpha = \frac{3.699}{\sqrt{\bar{\mu}/\bar{\epsilon}} \bar{g} b^2}. \quad (380)$$

The attenuation constant of a complete Clogston 2 with the same stack parameters $\bar{\mu}$ and $\bar{\epsilon}$ is, from (318),

$$\alpha = \frac{7.341}{\sqrt{\bar{\mu}/\bar{\epsilon}} \bar{g} b^2}, \quad (381)$$

so that the attenuation constant of the optimum loaded cable is only

about 0.504 times that of the optimum unloaded one. In this example, of course, we have said nothing about the effects of magnetic dissipation.

In the above work we have assumed that the electrical constants $\bar{\mu}$, $\bar{\epsilon}$, \bar{g} of the stacks and μ_0 , ϵ_0 of the main dielectric were all fixed quantities. We now consider the case in which the electrical constants of the conducting and insulating layers are given, but the fraction θ of conducting material in the stacks may be varied. We also suppose that the constants of the main dielectric are at our disposal, so that Clogston's condition may always be satisfied. When then is the optimum value of θ ?

If we express $\bar{\epsilon}$, $\bar{\mu}$, and \bar{g} in terms of the constants of the individual layers by equations (268) of Section VIII, we find that the expression for the attenuation constant of the principal mode in a Clogston line becomes

$$\alpha = \frac{\chi_1^2}{2\sqrt{\bar{\mu}/\bar{\epsilon}}\bar{g}} = \frac{\sqrt{\epsilon_2}\chi_1^2}{2\theta(1-\theta)^{\frac{1}{2}}[\theta\mu_1 + (1-\theta)\mu_2]^{\frac{1}{2}}g_1}, \quad (382)$$

where χ_1 is the lowest root of equation (348) for a plane line or equation (364) for a coaxial cable. We wish to minimize α as a function of θ .

If the conducting and insulating layers have different permeabilities ($\mu_1 \neq \mu_2$), then in the general partially filled line χ_1 depends on θ , through the factor $\bar{\mu}$ in equation (348) or (364), as well as on the geometric proportions of the line. In the limiting case of an extreme Clogston 1 line we found in Section IV, equation (145), that the optimum value of θ is

$$\theta = \frac{\mu_1 + (\mu_1^2 + 8\mu_1\mu_2)^{\frac{1}{2}}}{3\mu_1 + (\mu_1^2 + 8\mu_1\mu_2)^{\frac{1}{2}}}; \quad (383)$$

while in a Clogston 2 with no main dielectric, it turns out from (348) or (364) that χ_1 is independent of θ , and an elementary calculation shows that the value of θ which minimizes α is

$$\theta = \frac{3(\mu_1 - 2\mu_2) + (9\mu_1^2 - 4\mu_1\mu_2 + 4\mu_2^2)^{\frac{1}{2}}}{8(\mu_1 - \mu_2)}. \quad (384)$$

For the general partially filled line, however, there seems to be no simple expression for the optimum value of θ .

If the conducting and insulating layers have equal permeabilities, then the average permeability $\bar{\mu}$ ($= \mu_1 = \mu_2$) is independent of θ , and matters are much simpler. Since χ_1 is also independent of θ , the minimum value of α in equation (382) is achieved when

$$\theta = 2/3, \quad (385)$$

that is, when *the thickness of the conducting layers is twice the thickness of the insulating layers*. Thus the result obtained in Section IV for extreme Clogston 1 lines is shown to hold for Clogston lines with an arbitrary degree of filling, provided only that the permeabilities of the conducting and insulating layers are equal.

We emphasize that the preceding results apply only when the layers are infinitesimally thin. If the layers are of finite thickness, then the optimum value of θ will be less than that calculated for infinitesimally thin layers. The case of finite layers will be discussed in Section XI.

X. HIGHER MODES IN CLOGSTON LINES

We shall now investigate certain of the higher modes which are possible in Clogston-type transmission lines. As elsewhere in this paper, we shall restrict ourselves to modes having H_x , E_y , E_z or H_ϕ , E_ρ , E_z field components only, and for simplicity we shall assume stacks of infinitesimally thin laminae backed by high-impedance boundaries; but we shall place no restrictions on the relative thicknesses of the stacks and the main dielectric. We shall suppose, however, that the main dielectric always satisfies Clogston's condition. From physical considerations we anticipate the existence of higher modes of two types:

(1) In a partially filled Clogston line containing a finite thickness of main dielectric, there will be a group of modes very similar to the modes which can propagate between perfect conductors when the frequency is high enough to allow one or more field reversals in the space between the conductors. In a Clogston line these modes will have most of their field energy in the main dielectric, and for lack of a better term may be called "dielectric modes". They will all be cut off at sufficiently low frequencies, and for this reason are not likely to be of much engineering importance. The cutoff frequency of any particular dielectric mode is approximately inversely proportional to the thickness of the main dielectric, so that these modes cannot exist in a completely filled Clogston 2.

(2) There will also be a group of modes which are closely bound up with the laminated stacks, and which correspond to one or more current reversals in the stacks themselves; we shall call these the "stack modes".²¹ The stack modes will propagate down to zero frequency on either a partially or a completely filled Clogston line. They will have higher attenuation constants than the principal mode, but occasions may arise in which they are of considerable practical importance. We shall therefore consider these modes in some detail in what follows.

²¹ The stack modes in plane lines were discussed by Clogston in Reference 1, Sections IV-VI.

As we have seen in the preceding section, the even and odd modes in a plane Clogston line with infinitesimally thin laminae and high-impedance boundaries correspond respectively to the roots of equations (331) and (332), namely

$$\tanh \frac{1}{2} \sqrt{i\omega\bar{\epsilon}/\bar{g}} \Gamma_l b \tanh \Gamma_l s = -\frac{\bar{\mu}}{\mu_0} \sqrt{\frac{i\omega\bar{\epsilon}}{\bar{g}}}, \quad (386)$$

$$\coth \frac{1}{2} \sqrt{i\omega\bar{\epsilon}/\bar{g}} \Gamma_l b \tanh \Gamma_l s = -\frac{\bar{\mu}}{\mu_0} \sqrt{\frac{i\omega\bar{\epsilon}}{\bar{g}}}. \quad (387)$$

In either case the propagation constant γ is related to Γ_l by

$$\gamma^2 = -\omega^2 \bar{\mu} \bar{\epsilon} - (i\omega\bar{\epsilon}/\bar{g}) \Gamma_l^2, \quad (388)$$

and the field components are given by (333) and (334) for the even modes, or by (335) and (336) for the odd modes.

Our first observation relative to equations (386) and (387) is that the right-hand sides of these equations are extremely small compared to unity. Since the right-hand members are of the order of magnitude of $(\omega\bar{\epsilon}/\bar{g})^{\frac{1}{2}}$, at least one of the two factors on the left side of each equation must be of the order of $(\omega\bar{\epsilon}/\bar{g})^{\frac{1}{2}}$, which is still small compared to unity. If we consider the factors separately, there will be an infinite number of values of Γ_l for which each vanishes, since the hyperbolic tangent vanishes whenever its argument is equal to $m\pi i$, where m is any integer, and the hyperbolic cotangent vanishes whenever its argument equals $(m + \frac{1}{2})\pi i$. Since the coefficients of Γ_l in the two factors on the left side of either equation have different phase angles, we see that both factors cannot vanish simultaneously for any non-zero value of Γ_l . However as we have noted earlier the coefficient of Γ_l in the first factor is very much smaller than the coefficient of Γ_l in the second factor, and so in equation (386) both hyperbolic tangents may be small in the neighborhood of the first few non-zero roots of the second one. On the other hand the second hyperbolic tangent will not be small in the neighborhood of the non-zero roots of the first one; and in equation (387) the hyperbolic tangent and cotangent will never be small simultaneously. With these remarks in mind we shall proceed to a more detailed study of the various higher modes.

One group of modes is given to a good approximation by the condition that the first factor on the left side of equation (386) or (387) vanishes, that is,

$$\sqrt{i\omega\bar{\epsilon}/\bar{g}} \Gamma_l b \approx m\pi i, \quad (389)$$

where $m = 1, 2, 3, \dots$, and the even values of m correspond to the even

modes while the odd values correspond to the odd modes. In this section we shall exclude the case $m = 0$, which corresponds to the principal mode discussed in the preceding section. From (389) and equation (330) of Section IX, we get

$$\Gamma_{\ell} \approx \frac{m\pi}{b} \sqrt{\frac{i\bar{g}}{\omega\bar{\epsilon}}} = \frac{(1+i)m\pi}{\sqrt{2}b} \sqrt{\frac{\bar{g}}{\omega\bar{\epsilon}}}, \quad (390)$$

$$\kappa_0 \approx \frac{m\pi i}{b}. \quad (391)$$

The fields of the m th mode are given by substituting these quantities into (333) and (334) if m is even, or (335) and (336) if m is odd.

From equation (388), making use of Clogston's condition, the propagation constant of the m th mode of this family is given by

$$\gamma^2 \approx -\omega^2 \mu_0 \epsilon_0 + m^2 \pi^2 / b^2 = -4\pi^2 / \lambda_0^2 + m^2 \pi^2 / b^2, \quad (392)$$

where λ_0 is the wavelength of a free wave in the main dielectric at the operating frequency. To this approximation the values of γ are the same as the propagation constants of the family of modes (with H_x , E_y , and E_z field components only) which are possible in a dielectric slab of thickness b between perfectly conducting planes. The cutoff wavelength of the m th mode is

$$\lambda_c = 2b/m, \quad (393)$$

the propagation constant being real, to the present approximation, if $\lambda_0 > \lambda_c$ and pure imaginary if $\lambda_0 < \lambda_c$. We see that the cutoff frequency is inversely proportional to the width of the main dielectric, so that this family of modes is not possible in a completely filled Clogston 2.

It is worth noting that the effective skin depth of the stacks for the m th mode is, from (390),

$$\Delta = \frac{1}{\text{Re } \Gamma_{\ell}} = \frac{b}{m\pi} \sqrt{\frac{2\omega\bar{\epsilon}}{\bar{g}}} = \frac{\lambda_c}{\lambda_0} \sqrt{\frac{2}{\omega\bar{\mu}\bar{g}}}. \quad (394)$$

If the mode is just above cutoff, then Δ is of the order of magnitude of δ_1 ($= \sqrt{2/\omega\mu_1 g_1}$), but as ω increases indefinitely Δ also increases indefinitely, for the ideal stack of infinitesimally thin laminae. The physical explanation is simple: When the mode is near cutoff the phase velocity is very high, but as the frequency is increased the phase velocity approaches the velocity of a free wave in the main dielectric, for which the effective skin depth of the stacks was designed by Clogston's condi-

tion to be infinite. By increasing the $\mu_0\epsilon_0$ product of the main dielectric, it would be possible to make the effective skin thickness of a stack of infinitesimally thin layers infinite for any given mode at any single specified frequency, but at the moment this possibility appears of scarcely more than academic interest. Of course the practical limitation on effective skin depth at high frequencies is the finite thickness of the layers, a consideration which we do not take into account in the present section.

The attenuation constants of the dielectric modes, when these modes are above cutoff, may be calculated either by obtaining the small corrections to the values of Γ_l due to the fact that the right side of equation (386) or (387) is not rigorously zero, or by taking one-half the ratio of dissipated power per unit length to transmitted power. Either method gives for the m th mode, assuming the stack thickness s to be large compared to Δ ,

$$\alpha = \frac{m\pi}{b^2} \frac{\bar{\mu}}{\mu_0} \frac{\sqrt{2/\omega\bar{\mu}\bar{g}}}{\sqrt{1 - (\lambda_0/\lambda_c)^2}}. \quad (395)$$

Equation (395) assumes conducting layers very thin compared to the skin depth, a situation which may be difficult to achieve at frequencies high enough to permit the modes of this family to propagate.

Another family of modes which can exist on a parallel-plane Clogston line is given by the condition that the second factor on the left side of equation (386) or (387) shall be nearly equal to zero. As pointed out above the even modes present a slight complication; since the coefficient of Γ_l in the first hyperbolic tangent on the left side of (386) is very small, this factor may be comparable to or smaller than the term on the right side in the neighborhood of the first few roots of the equation, in which case the second hyperbolic tangent will not be small compared to unity at these roots. For all the modes in which we can conceivably be interested, however, $|\Gamma_l b|$ will be a small fraction of the very large number $2\sqrt{\bar{g}/\omega\bar{\epsilon}}$, and we may therefore replace the first hyperbolic tangent on the left side of (386) by its argument. Thus on making the usual substitution,

$$\Gamma_l^2 = -\chi^2, \quad \Gamma_l = i\chi, \quad (396)$$

we get for the even modes,

$$\chi s \tan \chi s = \frac{\bar{\mu}}{\mu_0} \frac{2s}{b}, \quad (397)$$

which is the same as equation (348) of the preceding section. On the

other hand, the odd modes of this family are all given approximately by

$$\tan \chi s = 0, \quad (398)$$

since the hyperbolic cotangent on the left side of (387) will never be small for the same value of Γ_t (or χ) as the hyperbolic tangent.

Equation (397) has an infinite number of positive real roots, which may be denoted by

$$\chi_1, \chi_3, \chi_5, \dots, \chi_{2p+1}, \dots, \quad (399)$$

and it is clear that

$$p\pi/s < \chi_{2p+1} \leq (p + \frac{1}{2})\pi/s. \quad (400)$$

If the thickness b of the main dielectric is not zero, so that the right side of equation (397) is finite, the higher roots χ_{2p+1} approach nearer and nearer to $p\pi/s$ as p increases; but if $b = 0$, then

$$\chi_{2p+1} = (p + \frac{1}{2})\pi/s = (2p + 1)\pi/a \quad (401)$$

for all p . The positive roots of equation (398) may be called

$$\chi_2, \chi_4, \chi_6, \dots, \chi_{2p}, \dots, \quad (402)$$

where

$$\chi_{2p} = p\pi/s \quad (403)$$

for all p ; and both sets of roots may be combined in the single sequence

$$\chi_1, \chi_2, \chi_3, \dots, \chi_p, \dots. \quad (404)$$

The advantages of designating the principal mode as the first rather than the zero-th mode seem to outweigh the minor disadvantage that in the sequence (404) the odd subscripts correspond to what we have been calling the even modes, and vice versa.

The attenuation and phase constants of the p th mode are obtained in terms of χ_p from (388) and (396). Under the usual assumption that the attenuation per radian is small, we have

$$\alpha = \frac{\chi_p^2}{2\sqrt{\bar{\mu}/\bar{\epsilon}} \bar{g}}, \quad (405)$$

$$\beta = \omega\sqrt{\bar{\mu}\bar{\epsilon}}, \quad (406)$$

which become, for the completely filled Clogston 2,

$$\alpha = \frac{p^2\pi^2}{2\sqrt{\bar{\mu}/\bar{\epsilon}} \bar{g}a^2}, \quad (407)$$

$$\beta = \omega\sqrt{\bar{\mu}\bar{\epsilon}}. \quad (408)$$

From (330) and (396) we have for the p th mode,

$$\Gamma_t = i\chi_p, \quad (409)$$

$$\kappa_0 = (-1 + i)\sqrt{\omega\bar{\epsilon}/2\bar{g}} \chi_p. \quad (410)$$

The fields may be obtained by substituting Γ_t and κ_0 into equations (333) and (334) when p is odd, or (335) and (336) when p is even. For the modes in which we are interested, that is, for sufficiently small values of p , we may replace $\text{sh } \kappa_0 y$ by $\kappa_0 y$ and $\text{ch } \kappa_0 y$ by unity when $|y| \leq \frac{1}{2}b$. Then for the modes with odd subscripts $2p + 1$ we have, in the main dielectric,

$$\begin{aligned} H_x &\approx H_0 e^{-\gamma_{2p+1}z}, \\ E_y &\approx -\sqrt{\frac{\mu_0}{\epsilon_0}} H_0 e^{-\gamma_{2p+1}z}, \\ E_z &\approx \frac{\mu_0 \chi_{2p+1}^2}{\bar{\mu}\bar{g}} H_0 e^{-\gamma_{2p+1}z}, \end{aligned} \quad (411)$$

for $-\frac{1}{2}b \leq y \leq \frac{1}{2}b$, while in the stacks,

$$\begin{aligned} H_x &\approx H_0 \frac{\sin \chi_{2p+1}(\frac{1}{2}a \mp y)}{\sin \chi_{2p+1}s} e^{-\gamma_{2p+1}z}, \\ \bar{E}_y &\approx -\sqrt{\frac{\bar{\mu}}{\bar{\epsilon}}} H_0 \frac{\sin \chi_{2p+1}(\frac{1}{2}a \mp y)}{\sin \chi_{2p+1}s} e^{-\gamma_{2p+1}z}, \\ E_z &\approx \pm \frac{\chi_{2p+1}}{\bar{g}} H_0 \frac{\cos \chi_{2p+1}(\frac{1}{2}a \mp y)}{\sin \chi_{2p+1}s} e^{-\gamma_{2p+1}z}, \end{aligned} \quad (412)$$

for $\frac{1}{2}b \leq |y| \leq \frac{1}{2}a$, where the upper signs refer to the upper stack and the lower signs to the lower stack. Of course the arbitrary amplitude factor H_0 need not be the same for different values of p . Similarly for the modes with even subscripts $2p$ the fields in the main dielectric are

$$\begin{aligned} H_x &\approx 0, \\ E_y &\approx 0, \\ E_z &\approx (-)^p \frac{p\pi}{\bar{g}s} H_0 e^{-\gamma_{2p}z}, \end{aligned} \quad (413)$$

for $-\frac{1}{2}b \leq y \leq \frac{1}{2}b$, while in the stacks,

$$\begin{aligned}
H_x &\approx \pm H_0 \sin \frac{p\pi(\frac{1}{2}a \mp y)}{s} e^{-\gamma_{2p}z}, \\
\bar{E}_y &\approx \mp \sqrt{\frac{\bar{\mu}}{\bar{\epsilon}}} H_0 \sin \frac{p\pi(\frac{1}{2}a \mp y)}{s} e^{-\gamma_{2p}z}, \\
E_z &\approx \frac{p\pi}{\bar{g}s} H_0 \cos \frac{p\pi(\frac{1}{2}a \mp y)}{s} e^{-\gamma_{2p}z},
\end{aligned} \tag{414}$$

for $\frac{1}{2}b \leq |y| \leq \frac{1}{2}a$, and again the upper signs refer to the upper stack and the lower signs to the lower stack.

In a complete Clogston 2 the expressions for the fields simplify a good deal. For the modes with odd subscripts $2p + 1$ the fields are, for $-\frac{1}{2}a \leq y \leq \frac{1}{2}a$,

$$\begin{aligned}
H_x &\approx H_0 \cos \frac{(2p+1)\pi y}{a} e^{-\gamma_{2p+1}z}, \\
\bar{E}_y &\approx -\sqrt{\frac{\bar{\mu}}{\bar{\epsilon}}} H_0 \cos \frac{(2p+1)\pi y}{a} e^{-\gamma_{2p+1}z}, \\
E_z &\approx \frac{(2p+1)\pi}{\bar{g}a} H_0 \sin \frac{(2p+1)\pi y}{a} e^{-\gamma_{2p+1}z},
\end{aligned} \tag{415}$$

while for the modes with even subscripts $2p$,

$$\begin{aligned}
H_x &\approx H_0 \sin \frac{2p\pi y}{a} e^{-\gamma_{2p}z}, \\
\bar{E}_y &\approx -\sqrt{\frac{\bar{\mu}}{\bar{\epsilon}}} H_0 \sin \frac{2p\pi y}{a} e^{-\gamma_{2p}z}, \\
E_z &\approx -\frac{2p\pi}{\bar{g}a} H_0 \cos \frac{2p\pi y}{a} e^{-\gamma_{2p}z}.
\end{aligned} \tag{416}$$

The fields of the higher modes in a plane Clogston 2 are simply related to the fields of the principal mode shown in Fig. 13(d). The fields of the p th mode may be obtained conceptually by stacking up p "layers", each of thickness a/p , the fields in each layer being identical with the fields of the principal mode except for the scale reduction and a phase difference of 180° between adjacent layers. Equation (407) shows that the attenuation constant of the p th mode in a plane Clogston 2 with infinitesimally thin laminae and high-impedance walls is just p^2 times the attenuation constant of the principal mode.

It may be observed that if we are considering a partially filled plane Clogston line with $b > 0$, then the propagation constants of the $2p$ th

and the $(2p + 1)$ st stack modes will be nearly the same for sufficiently large values of p (how large depends on the ratio of stack thickness to main dielectric thickness). Except for differences in sign, the fields in the stacks will also be the same up to second order differences which our approximations do not show. The physical interpretation is that for a thick enough main dielectric and/or sufficiently large values of p , the fields are confined almost entirely to the two stacks, being relatively small in the main dielectric, while the stacks act like a pair of almost independent Clogston 2 lines each of thickness s and carrying a particular Clogston 2 higher mode.

Figs. 16 and 17 show field plots for the second and third stack modes (i.e., the first and second higher modes) in the same four plane Clogston lines that were used to exhibit the behavior of the principal mode in Fig. 13. Note however that these plots are not normalized and that the horizontal scales on the figures are not all the same. The following table gives $\chi_2 s$ and $\chi_3 s$ as functions of the fraction $s/\frac{1}{2}a$ of the total space filled by the stacks, and also the quantities $(\chi_2 a/\pi)^2$ and $(\chi_3 a/\pi)^2$, which are just the ratios of the attenuation constants of

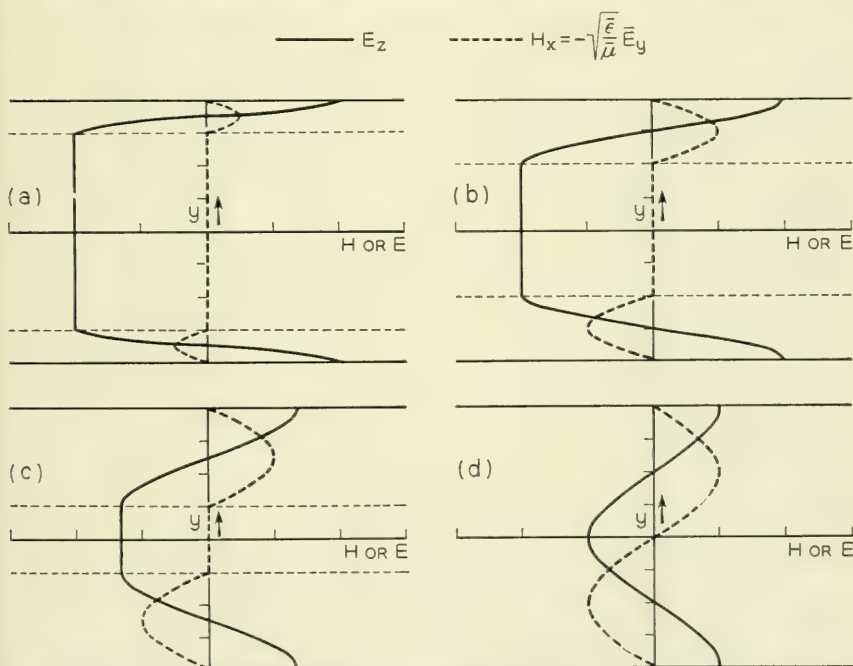


Fig. 16—Fields of second stack mode in partially and completely filled plane Clogston lines with $\mu_0 = \bar{\mu}$, $\epsilon_0 = \bar{\epsilon}$.

these modes to the attenuation constant of the principal mode in a completely filled Clogston 2.

$s/\frac{1}{2}a$	χ^{2S}	χ^{3S}	$(\chi^{2a}/\pi)^2$	$(\chi^{3a}/\pi)^2$
$\frac{1}{4}$	π	3.244	64.0	68.2
$\frac{1}{2}$	π	3.426	16.0	19.0
$\frac{3}{4}$	π	3.809	7.1	10.5
1	π	4.712	4.0	9.0

These results may be compared with those given in Section IX for the principal mode in plane Clogston lines having the same proportions.

Turning now to the cylindrical geometry, we remember that all the circular transverse magnetic modes in a coaxial Clogston line with infinitesimally thin laminae and high-impedance boundaries are given by the roots of equation (339) of Section IX, namely

$$\frac{\kappa_0 K_0(\kappa_0 \rho_1) + i \omega \epsilon_0 Z_1 K_1(\kappa_0 \rho_1)}{\kappa_0 I_0(\kappa_0 \rho_1) - i \omega \epsilon_0 Z_1 I_1(\kappa_0 \rho_1)} = \frac{\kappa_0 K_0(\kappa_0 \rho_2) - i \omega \epsilon_0 Z_2 K_1(\kappa_0 \rho_2)}{\kappa_0 I_0(\kappa_0 \rho_2) + i \omega \epsilon_0 Z_2 I_1(\kappa_0 \rho_2)}, \quad (417)$$

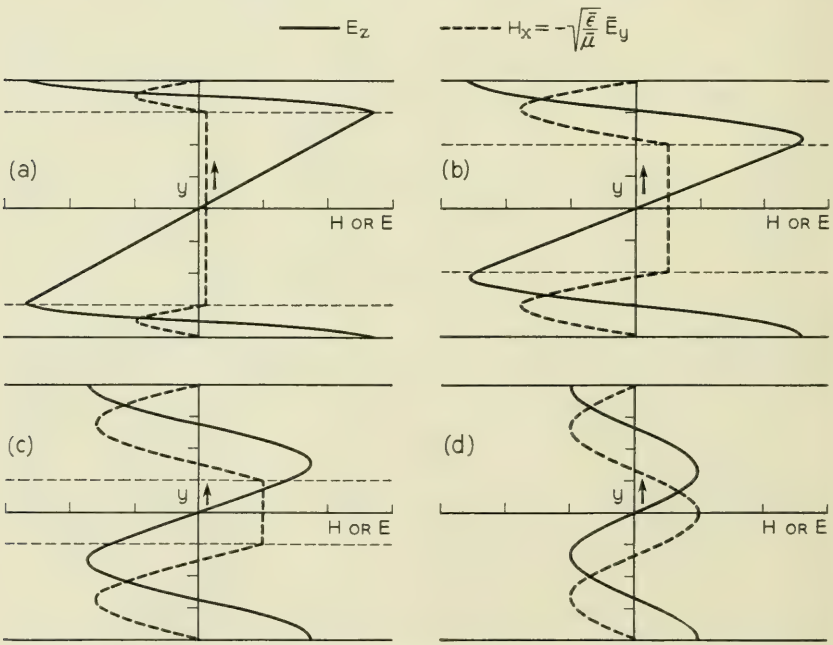


Fig. 17—Fields of third stack mode in partially and completely filled plane Clogston lines with $\mu_0 = \bar{\mu}$, $\epsilon_0 = \bar{\epsilon}$.

where

$$Z_1 = \frac{\Gamma_\ell K_0(\Gamma_\ell \rho_1) I_1(\Gamma_\ell a) + K_1(\Gamma_\ell a) I_0(\Gamma_\ell \rho_1)}{\bar{g} \bar{K}_1(\Gamma_\ell a) I_1(\Gamma_\ell \rho_1) - K_1(\Gamma_\ell \rho_1) I_1(\Gamma_\ell a)}, \quad (418)$$

$$Z_2 = \frac{\Gamma_\ell K_0(\Gamma_\ell \rho_2) I_1(\Gamma_\ell b) + K_1(\Gamma_\ell b) I_0(\Gamma_\ell \rho_2)}{\bar{g} \bar{K}_1(\Gamma_\ell \rho_2) I_1(\Gamma_\ell b) - K_1(\Gamma_\ell b) I_1(\Gamma_\ell \rho_2)}. \quad (419)$$

Physically it is clear that the modes of the coaxial cable must be of the same general types as the modes of the parallel-plane line, and so in seeking the roots of equation (417) we shall be guided by the results which we have already found for the plane structure.

The dielectric modes in the cable may be located, to a first approximation, by setting Z_1 and Z_2 equal to zero, whence (417) becomes

$$\frac{K_0(\kappa_0 \rho_1)}{I_0(\kappa_0 \rho_1)} = \frac{K_0(\kappa_0 \rho_2)}{I_0(\kappa_0 \rho_2)}. \quad (420)$$

The substitution

$$\kappa_0^2 = -h^2, \quad \kappa_0 = ih, \quad (421)$$

transforms (420) into

$$J_0(h\rho_1)N_0(h\rho_2) - J_0(h\rho_2)N_0(h\rho_1) = 0. \quad (422)$$

Equation (422) has an infinite number of real roots $h_1, h_2, h_3, \dots, h_m, \dots$, of which the m th one may be written in the form²²

$$h_m = \frac{m\pi F_m(\rho_1/\rho_2)}{\rho_2 - \rho_1}, \quad (423)$$

where $F_m(\rho_1/\rho_2)$ is a function which increases from slightly less than unity at $\rho_1/\rho_2 = 0$ to unity at $\rho_1/\rho_2 = 1$. From equations (330) and (423) we have, approximately,

$$\Gamma_\ell = h_m \sqrt{\frac{i\bar{g}}{\omega\bar{\epsilon}}} = \frac{(1+i)m\pi F_m(\rho_1/\rho_2)}{\sqrt{2}(\rho_2 - \rho_1)} \sqrt{\frac{\bar{g}}{\omega\bar{\epsilon}}}, \quad (424)$$

$$\kappa_0 = ih_m = \frac{im\pi F_m(\rho_1/\rho_2)}{\rho_2 - \rho_1}, \quad (425)$$

and the fields of the m th mode are given by substituting these expressions into equations (340) to (344) of the preceding section.

From equation (388) the propagation constant of the m th dielectric

²² Reference 18, pp. 204-206. What we call $m\pi F_m(\rho_1/\rho_2)$ is tabulated by Jahneke and Emde, pp. 205-206, as $(k-1)x_0^{(m)}$, where $k = \rho_2/\rho_1$.

mode is defined by

$$\gamma^2 = -\omega^2 \mu_0 \epsilon_0 + h_m^2 = -4\pi^2/\lambda_0^2 + h_m^2 \quad (426)$$

to the present approximation, and the cutoff wavelength is

$$\lambda_c = \frac{2\pi}{h_m} = \frac{2(\rho_2 - \rho_1)}{mF_m(\rho_1/\rho_2)}, \quad (427)$$

which tends to zero with the thickness $\rho_2 - \rho_1$ of the main dielectric.

As in the parallel-plane case, when the m th dielectric mode is just able to propagate the effective skin depth in the stacks is of the order of δ_1 , and the stack impedances are approximately

$$Z_1 = Z_2 = K = \Gamma_t/\bar{g}, \quad (428)$$

under the present assumption of infinitesimally thin laminae. The power dissipated in the stacks and the corresponding attenuation constant may be calculated by a straightforward procedure if desired.

Before leaving the subject of higher dielectric modes in a Clogston cable, we should point out that although we have mentioned only the transverse magnetic modes with circular symmetry, in reality there exist a double infinity of both transverse magnetic and transverse electric higher modes. These modes are discussed in textbooks²³ for coaxial lines bounded by perfect conductors, and they will propagate, with minor changes due to wall losses, in either ordinary or Clogston-type coaxial cables if the frequency is high enough. At ordinary engineering frequencies, however, the higher modes contribute only to the local fields excited at discontinuities, and are therefore not of any great practical importance.

To find the stack modes in a Clogston cable we assume, subject to a posteriori verification, that in the main dielectric we shall have $|\kappa_0\rho| \ll 1$ for all the modes of interest. Then if we set $\Gamma_t = i\chi$, equation (417) reduces, as in Section IX, to

$$\begin{aligned} & \frac{1}{\chi\rho_1} \frac{J_1(\chi a)N_0(\chi\rho_1) - N_1(\chi a)J_0(\chi\rho_1)}{J_1(\chi a)N_1(\chi\rho_1) - N_1(\chi a)J_1(\chi\rho_1)} \\ & + \frac{1}{\chi\rho_2} \frac{J_1(\chi b)N_0(\chi\rho_2) - N_1(\chi b)J_0(\chi\rho_2)}{J_1(\chi a)N_1(\chi b) - N_1(\chi\rho_2)J_1(\chi b)} = \frac{\mu_0}{\bar{\mu}} \log \frac{\rho_2}{\rho_1}, \end{aligned} \quad (429)$$

which is the same as equation (364). Equation (429) has an infinite

²³ A good account is given by N. Marcuvitz, *Waveguide Handbook*, M. I. T. Rad. Lab. Series, **10**, McGraw-Hill, New York, 1951, pp. 72-80.

number of real roots,

$$\chi_1, \chi_2, \chi_3, \dots, \chi_p, \dots, \quad (430)$$

of which χ_1 corresponds to the principal mode and χ_2, χ_3, \dots , to the higher stack modes. The χ 's are the eigenvalues of the system of equations (369), and as such may be located approximately with a differential analyzer, or as accurately as desired by numerical solution of equation (429). The attenuation and phase constants of the p th mode are

$$\alpha = \frac{\chi_p^2}{2\sqrt{\bar{\mu}/\bar{\epsilon}} \bar{g}}, \quad (431)$$

$$\beta = \omega\sqrt{\bar{\mu}\bar{\epsilon}}, \quad (432)$$

provided that the attenuation per radian is small, i.e., that p is not too large. The fields are given by writing χ_p for χ_1 and γ_p for γ in equations (371) to (373) of Section IX.

For a Clogston 2 with no main dielectric we can set $\rho_1 = \rho_2$ in equation (429) and obtain the much simpler form

$$J_1(\chi a)N_1(\chi b) - J_1(\chi b)N_1(\chi a) = 0. \quad (433)$$

The p th root of (433) may be written²⁴

$$\chi_p = \frac{p\pi f_p(a/b)}{b - a}, \quad (434)$$

where the functions $f_p(a/b)$ have values slightly greater than unity when $a/b = 0$, and decrease monotonically toward 1 as a/b approaches unity. The attenuation and phase constants of the p th mode in a Clogston 2 are given by

$$\alpha = \frac{p^2 \pi^2 f_p^2(a/b)}{2\sqrt{\bar{\mu}/\bar{\epsilon}} \bar{g}(b - a)^2}, \quad (435)$$

$$\beta = \omega\sqrt{\bar{\mu}\bar{\epsilon}}, \quad (436)$$

provided that p is not too large. The attenuation constant of the p th mode is thus approximately p^2 times the attenuation constant of the principal mode, the approximation being better the closer the ratio a/b is to unity. The fields of the p th mode may be obtained by writing χ_p for χ_1 and γ_p for γ in equations (302) of Section VIII, or equations (311) if $a = 0$. Qualitatively these fields are very similar to the fields of

²⁴ Reference 18, pp. 204-206. What we call $p\pi f_p(a/b)$ is tabulated by Jahnke and Emde as $(k - 1)x_1^{(p)}$, where $k = b/a$.

the p th mode in a parallel-plane Clogston 2, with the same number of field maxima and field reversals for a given value of p , though of course the spacings and amplitudes of the field maxima are not all equal in the coaxial cable.

As numerical examples we have plotted in Figs. 18 and 19 the fields of the second and third stack modes (i.e., the first and second higher modes) in the same two Clogston cables which were used to show the principal mode in Fig. 14. The horizontal scales on these figures are arbitrary and have no relation to one another. Figs. 18(a) and 19(a) represent a partially filled cable with the same dimensions, namely $a = 0.084b$, $\rho_1 = 0.415b$, and $\rho_2 = 0.831b$, as in Fig. 14(a), while Figs. 18(b) and 19(b) represent a completely filled cable, as in Fig. 14(b). The following table shows, as a function of the filling ratio $(s_1 + s_2)/b$, the quantity $(\chi_p b/3.83)^2$ for $p = 1, 2, 3$; this quantity is just the ratio of the attenuation constant of the given mode to the attenuation constant of the principal mode in a completely filled Clogston 2.

$(s_1 + s_2)/b$	$(\chi_1 b/3.83)^2$	$(\chi_2 b/3.83)^2$	$(\chi_3 b/3.83)^2$
0.5	1.234	8.369	24.273
1.0	1.000	3.352	7.050

We note that although the proportions of the partially filled cable were found in Section IX to be optimum, in the sense of minimizing the attenuation constant, for the principal mode in a cable with filling ratio 0.5, there is no reason to believe that the same proportions will be optimum for the second and third modes with the same filling ratio.

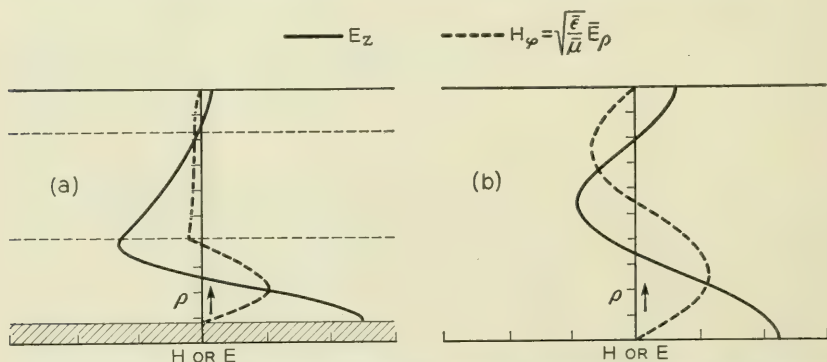


Fig. 18—Fields of second stack mode in partially and completely filled coaxial Clogston lines with $\mu_0 = \bar{\mu}$, $\epsilon_0 = \bar{\epsilon}$.

XI. EFFECT OF FINITE LAMINA THICKNESS. FREQUENCY DEPENDENCE OF ATTENUATION IN CLOGSTON 2 LINES

We shall now study Clogston 2 lines with laminae of finite thickness, and shall investigate the important practical question of how the propagation constant varies with frequency in such lines. Much of the analysis of the present section will deal with parallel-plane structures, but we may be confident that the results will also give at least a good qualitative estimate of the behavior of coaxial cables.

The notation for the plane Clogston 2, shown in Fig. 10, is the same as before, except that we now assume the thicknesses of the individual conducting and insulating layers to be t_1 and t_2 respectively. For definiteness we shall suppose that there are $2n$ conducting layers and $2n$ insulating layers in the whole stack, with a conducting layer next to the lower sheath and an insulating layer next to the upper sheath, though the precise arrangement is of no real importance if the number of layers is large. The total thickness a of the stack is $2n(t_1 + t_2)$, and the fraction of conducting material will as usual be called θ .

The boundary conditions for any mode (having H_x , E_y , and E_z field components only) require that the sum of the impedances looking in opposite directions normal to any plane $y = \text{constant}$ be zero. If we match impedances at the lower sheath $y = -\frac{1}{2}a$ and use equation (65) of Section III for the impedance looking into the stack, we have

$$\frac{\frac{1}{2}Z_n(\gamma)(K_1e^{2n\Gamma} + K_2e^{-2n\Gamma}) + K_1K_2 \operatorname{sh} 2n\Gamma}{Z_n(\gamma) \operatorname{sh} 2n\Gamma + \frac{1}{2}(K_1e^{-2n\Gamma} + K_2e^{2n\Gamma})} + Z_n(\gamma) = 0, \quad (437)$$

where Γ , K_1 , and K_2 are given by equations (61) and (63). If equation

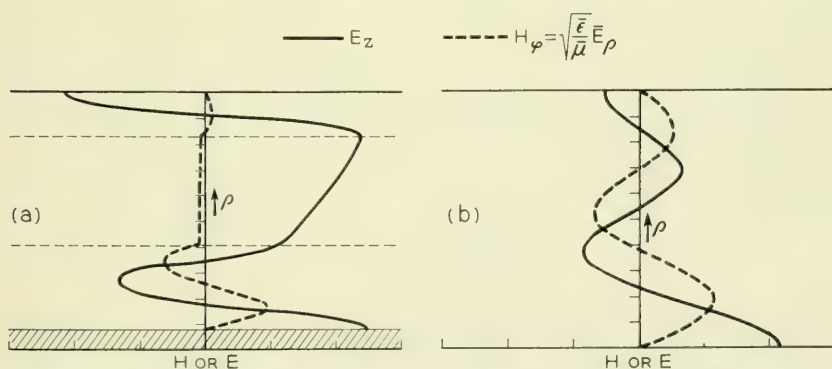


Fig. 19—Fields of third stack mode in partially and completely filled coaxial Clogston lines with $\mu_0 = \bar{\mu}$, $\epsilon_0 = \bar{\epsilon}$.

(437) is solved for $e^{4n\Gamma}$, it takes the form

$$e^{4n\Gamma} = \frac{[Z_n(\gamma) - K_1][Z_n(\gamma) - K_2]}{[Z_n(\gamma) + K_1][Z_n(\gamma) + K_2]}. \quad (438)$$

In order to simplify the general expressions for Γ , K_1 , and K_2 , we make the following approximations:

(i) We neglect $\omega\epsilon/g_1$ compared to unity, where ϵ represents the dielectric constant of either the conducting or the insulating layers. As we have said before, this is an exceedingly good approximation at all engineering frequencies.

(ii) We neglect γ^2/σ_1^2 ($= \gamma^2/i\omega\mu_1g_1$) compared to unity. It turns out that not only is this approximation valid in the frequency range of greatest interest, where γ is approximately equal to $i\omega\sqrt{\bar{\mu}\bar{\epsilon}}$, but also it is valid all the way down to zero frequency, so that in the present section we can easily derive results for the complete frequency range down to dc. So long as $\gamma^2/\sigma_1^2 \ll 1$, we have from equations (56) of Section III,

$$\kappa_1 \approx \sigma_1, \quad \eta_{1y} \approx \eta_1. \quad (439)$$

(iii) We suppose that the thickness t_2 of the layers of insulation is so small that we may replace $\text{sh } \kappa_2 t_2$ by $\kappa_2 t_2$ and $\text{ch } \kappa_2 t_2$ by unity. These approximations will be amply justified if t_2 is not greater than a few times the skin depth in the conducting layers at the highest operating frequency.

The foregoing approximations lead to results identical with equations (86) and (87) of Section III, namely

$$\text{ch } \Gamma = \frac{\eta_{2y}\kappa_2 t_2}{2\eta_{1y}} \text{sh } \kappa_1 t_1 + \text{ch } \kappa_1 t_1, \quad (440)$$

and

$$\begin{aligned} K_1 &= -\frac{1}{2}\eta_{2y}\kappa_2 t_2 + \sqrt{\left(\frac{1}{2}\eta_{2y}\kappa_2 t_2\right)^2 + \eta_{1y}\eta_{2y}\kappa_2 t_2 \coth \kappa_1 t_1 + \eta_{1y}^2}, \\ K_2 &= +\frac{1}{2}\eta_{2y}\kappa_2 t_2 + \sqrt{\left(\frac{1}{2}\eta_{2y}\kappa_2 t_2\right)^2 + \eta_{1y}\eta_{2y}\kappa_2 t_2 \coth \kappa_1 t_1 + \eta_{1y}^2}. \end{aligned} \quad (441)$$

To simplify the notation, we introduce the symbol q defined by

$$\begin{aligned} q &= -\frac{\eta_{2y}\kappa_2 t_2}{\eta_1\sigma_1 t_1} \\ &= -\frac{(1-\theta)\mu_2}{\theta\mu_1} \left[1 + \frac{\gamma^2}{\omega^2\mu_2\epsilon_2} \right] \\ &= 1 - \frac{\bar{\mu}}{\theta\mu_1} \left[1 + \frac{\gamma^2}{\omega^2\bar{\mu}\bar{\epsilon}} \right], \end{aligned} \quad (442)$$

where $\bar{\epsilon}$ and $\bar{\mu}$ are given by equations (268) of Section VIII. The propagation constant γ is thus related to q by

$$\begin{aligned}\gamma &= i\omega\sqrt{\mu_2\epsilon_2}\left[1 + \frac{\theta\mu_1q}{(1-\theta)\mu_2}\right]^{\frac{1}{2}} \\ &= i\omega\sqrt{\bar{\mu}\bar{\epsilon}}\left[1 + \frac{\theta\mu_1(q-1)}{\bar{\mu}}\right]^{\frac{1}{2}}.\end{aligned}\quad (443)$$

In terms of q and the electrical thickness parameter Θ used in Part I, namely

$$\Theta = \sigma_1 t_1 = (1+i)t_1/\delta_1 \approx \kappa_1 t_1, \quad (444)$$

equations (440) and (441) become, approximately,

$$\text{ch } \Gamma = \text{ch } \Theta - \frac{1}{2}q\Theta \text{ sh } \Theta, \quad (445)$$

and

$$\begin{aligned}K_1 &= \frac{\Theta}{g_1 t_1} \left[\frac{1}{2}q\Theta + \sqrt{\frac{1}{4}q^2\Theta^2 - q\Theta \coth \Theta + 1} \right], \\ K_2 &= \frac{\Theta}{g_1 t_1} \left[-\frac{1}{2}q\Theta + \sqrt{\frac{1}{4}q^2\Theta^2 - q\Theta \coth \Theta + 1} \right].\end{aligned}\quad (446)$$

In the general case when the sheath impedance $Z_n(\gamma)$ is a given function of γ , we substitute the expressions for K_1 and K_2 into equation (438), namely

$$e^{4n\Gamma} = \frac{Z_n^2(\gamma) - (K_1 + K_2)Z_n(\gamma) + K_1K_2}{Z_n^2(\gamma) + (K_1 + K_2)Z_n(\gamma) + K_1K_2}, \quad (447)$$

and then determine γ for each mode by simultaneous numerical solution of equations (443), (445), and (447). At least as a first approximation we may neglect the total current in either sheath compared to the one-way current in the stack; to this approximation $Z_n(\gamma)$ is effectively infinite and (447) becomes

$$e^{4n\Gamma} = 1. \quad (448)$$

The non-zero roots of this equation are

$$\Gamma = ip\pi/2n, \quad p = 1, 2, 3, \dots, \quad (449)$$

where $p = 1$ corresponds to the principal mode and the higher values of p to the higher modes discussed in Section X. (We would get nothing new by including negative values of p .) The quantities q and γ for each mode are then given by equations (445) and (443) respectively. If we

wish to take the finite value of $Z_n(\gamma)$ into account, we may calculate K_1 and K_2 from (446) and then obtain a second approximation to Γ from (447); and this process may be repeated as often as desired. From the experience gained in treating a particular example we feel that the method of successive approximations is entirely feasible, but it does involve a considerable amount of numerical work.

In the calculation just described we have to choose the correct sign of the complex square root occurring in the expressions for K_1 and K_2 . Without attempting to give a complete discussion of this point here, we observe that it may be shown that

$$\text{sh } \Gamma = \text{sh } \Theta \sqrt{\frac{1}{4}q^2\Theta^2 - q\Theta \coth \Theta + 1}. \quad (450)$$

Under ordinary circumstances Γ will be a small complex number with a phase angle of about 90° , and Θ will be a small complex number with a phase angle of 45° . Hence the phase angle of the square root may be expected to be in the neighborhood of 45° rather than 225° .

In the remainder of this section we shall restrict ourselves to the case of high-impedance sheaths, so that the values of Γ are given to a sufficiently good approximation by equation (449). We shall discuss the principal mode and the higher modes concurrently, but shall assume throughout that the mode number p is small compared to n . From (445) and (449), the value of q for the p th mode is

$$q = \frac{2 \left(\text{ch } \Theta - \cos \frac{p\pi}{2n} \right)}{\Theta \text{ sh } \Theta}, \quad (451)$$

and the propagation constant γ is obtained by substituting this value of q into equation (443).

We shall now discuss the variation of the propagation constant of a plane Clogston 2 line with frequency over the full frequency range from zero to very high frequencies.²⁵ To do this we shall derive approximate expressions for the propagation constant at what may be called, roughly, very low frequencies, low frequencies, high frequencies, and very high frequencies. It will appear presently that the limits of these various frequency ranges depend among other things on the dimensions of the laminated transmission line and the thicknesses of the individual layers, and that the frequency range of greatest engineering importance is what we have called simply "low frequencies".

From equation (444) we have

²⁵ In this connection see also Reference 1, Appendices A and B, pp. 525-529.

$$\Theta = (1 + i)t_1\sqrt{\omega\mu_1g_1/2} = (1 + i)t_1\sqrt{\pi\mu_1g_1f}, \quad (452)$$

which is proportional to the square root of frequency. For small Θ equation (451) may be written

$$\begin{aligned} q &= \left[2(\cosh \Theta - 1) + 4 \sin^2 \frac{p\pi}{4n} \right] \frac{\operatorname{csch} \Theta}{\Theta} \\ &= \left[4 \sin^2 \frac{p\pi}{4n} \right] \frac{1}{\Theta^2} + \left[1 - \frac{2}{3} \sin^2 \frac{p\pi}{4n} \right] \\ &\quad - \frac{1}{12} \left[1 - \frac{14}{15} \sin^2 \frac{p\pi}{4n} \right] \Theta^2 + \dots, \end{aligned} \quad (453)$$

on expanding the right side in powers of Θ by Dwight 657.2 and 657.8. If we replace $\sin p\pi/4n$ by $p\pi/4n$ and neglect the square of this quantity in comparison with unity, (453) becomes

$$q \approx \frac{p^2\pi^2}{4n^2} \frac{1}{\Theta^2} + 1 - \frac{\Theta^2}{12} + \dots. \quad (454)$$

Introducing this expression for q into equation (443) and substituting for Θ from (452), we get for the propagation constant,

$$\gamma = i\omega\sqrt{\bar{\mu}\bar{\epsilon}} \left[1 - \frac{i\theta\mu_1}{\bar{\mu}} \left(\frac{p^2\pi^2}{4n^2\omega\mu_1g_1t_1^2} + \frac{\omega\mu_1g_1t_1^2}{12} \right) \right]^{\frac{1}{2}}. \quad (455)$$

As the frequency approaches zero in a Clogston 2 line of finite thickness, the term in $1/\omega$ dominates the other terms in square brackets in equation (455). Thus at very low frequencies the attenuation and phase constants of the p th mode are given approximately by

$$\alpha = \frac{p\pi\theta}{2T_1} \sqrt{\frac{\omega\bar{\epsilon}}{2\bar{g}}} = \frac{p\pi}{a} \sqrt{\frac{\pi\bar{\epsilon}f}{\bar{g}}}, \quad (456)$$

$$\beta = \frac{p\pi\theta}{2T_1} \sqrt{\frac{\omega\bar{\epsilon}}{2\bar{g}}} = \frac{p\pi}{a} \sqrt{\frac{\pi\bar{\epsilon}f}{\bar{g}}}, \quad (457)$$

where $2T_1$ ($= 2nt_1$) is the total thickness of conducting material in the stack of thickness a . To this approximation the attenuation and phase constants are equal, and are proportional to the square root of frequency. We note that at very low frequencies,

$$\frac{\gamma^2}{\sigma_1^2} = \frac{2ip^2\pi^2\theta^2\omega\bar{\epsilon}}{8T_1^2\bar{g}} \cdot \frac{1}{i\omega\mu_1g_1} = \frac{\theta p^2\pi^2\bar{\epsilon}/\mu_1}{a^2\bar{g}^2}, \quad (458)$$

which will be very small compared to unity for lines of all reasonable dimensions, in agreement with our assumption (ii) above.

As the frequency is increased there will be a range in which the terms in parentheses in equation (455) are small compared to unity, so that the square root may be expanded by the binomial theorem. This gives

$$\alpha = \frac{p^2 \pi^2}{2\sqrt{\bar{\mu}/\bar{\epsilon}} \bar{g} a^2} + \frac{\omega^2 \mu_1^2 \bar{g} t_1^2}{24\sqrt{\bar{\mu}/\bar{\epsilon}}}, \quad (459)$$

$$\beta = i\omega\sqrt{\bar{\mu}\bar{\epsilon}}. \quad (460)$$

If the line is of finite total thickness a and the frequency is so low or the laminae are so thin that the first term on the right side of (459) is large compared to the second, we have approximately

$$\alpha = \frac{p^2 \pi^2}{2\sqrt{\bar{\mu}/\bar{\epsilon}} \bar{g} a^2}. \quad (461)$$

This is the frequency-independent attenuation constant that we found in Section X, equation (407), for the p th mode in a plane Clogston 2 with infinitesimally thin laminae. We shall call the range over which the attenuation is essentially flat the "low-frequency" range. On the other hand, if the laminae are of finite thickness the second term on the right side of (459) ultimately becomes dominant, and the attenuation constant is then given approximately by

$$\alpha = \frac{\omega^2 \mu_1^2 \bar{g} t_1^2}{24\sqrt{\bar{\mu}/\bar{\epsilon}}} = \frac{\pi^2 \mu_1^2 \bar{g} t_1^2 f^2}{6\sqrt{\bar{\mu}/\bar{\epsilon}}}. \quad (462)$$

This is also the attenuation constant of a plane wave in an unbounded laminated medium (except at very high frequencies), as may be seen by letting the stack thickness a tend to infinity in equation (459). By "high frequencies" we shall mean the frequency range in which the attenuation constant is approximately proportional to f^2 .

Finally at very high frequencies when $|\Theta| \gg 1$, we have from (451),

$$q \approx 2/\Theta, \quad (463)$$

and so from (443),

$$\gamma = i\omega\sqrt{\mu_2\epsilon_2} \left[1 + \frac{2\theta\mu_1}{(1-\theta)\mu_2\Theta} \right]^{\frac{1}{2}}. \quad (464)$$

Expanding by the binomial theorem and substituting for Θ from (452), we get after a little rearrangement,

$$\alpha = \frac{1}{\sqrt{\mu_2/\epsilon_2} t_2 g_1 \delta_1} = \frac{1}{\sqrt{\mu_2/\epsilon_2} t_2} \sqrt{\frac{\pi \mu_1 f}{g_1}}, \quad (465)$$

$$\beta = \omega \sqrt{\mu_2 \epsilon_2} + \frac{1}{\sqrt{\mu_2/\epsilon_2} t_2 g_1 \delta_1}. \quad (466)$$

Comparing these expressions with equations (25) and (26) of Section II, we see that they correspond to waves in parallel-plane transmission lines of width t_2 , bounded by electrically thick solid conductors. We shall call this range, in which α is proportional to the square root of frequency, the "very high frequency" range. At these frequencies the propagation constant is the same in a Clogston 2 line of finite thickness as in an infinite laminated medium.

In order to describe the various frequency ranges more precisely, we shall define the three critical frequencies f'_1 , f'_2 , and f'_3 to be the frequencies at which the approximate expressions for the attenuation constants in two adjacent frequency ranges are equal. These frequencies are closely related to the critical frequencies which we defined in equations (178) of Section V, when we were discussing the surface impedance of a plane stack of finite layers. For a stack containing a total thickness T_1 of conducting material, we recall that the critical frequencies were f_1 , where $T_1 = \delta_1$; f_2 , where $T_1 = T_\Delta$; and f_3 , where $t_1 = \sqrt{3} \delta_1$. For the p th mode in a Clogston 2 containing a total thickness $2T_1$ of conducting material, the frequencies turn out to be

$$\begin{aligned} f'_1 &= \frac{p^2 \pi \theta}{16 \bar{\mu} g_1 T_1^2} = \frac{p^2 \pi^2 \theta \mu_1}{16 \bar{\mu}} f_1, \\ f'_2 &= \frac{\sqrt{3} p}{2 \mu_1 g_1 t_1 T_1} = \frac{p \pi}{2} f_2, \\ f'_3 &= \left[\frac{36 \bar{\mu}}{(1 - \theta) \mu_2} \right]^{\frac{1}{3}} \frac{1}{\pi \mu_1 g_1 t_1^2} = \left[\frac{4 \bar{\mu} \epsilon}{3 \mu_2 \epsilon_2} \right]^{\frac{1}{3}} f_3, \end{aligned} \quad (467)$$

where of course $p = 1$ for the principal mode. Thus the attenuation constant is given approximately by (456) for $0 \leq f \leq f'_1$, by (461) for $f'_1 \leq f \leq f'_2$, by (462) for $f'_2 \leq f \leq f'_3$, and by (465) for $f \geq f'_3$.

If we plot the foregoing approximate expressions for the attenuation constant against frequency on log-log paper, we can get a good idea of the variation of the attenuation of a Clogston 2 over the entire frequency range. Both the approximate expressions and the exact results for a particular numerical case are plotted in Fig. 20, for a Clogston 2 of finite thickness and also for an infinite laminated medium. The actual

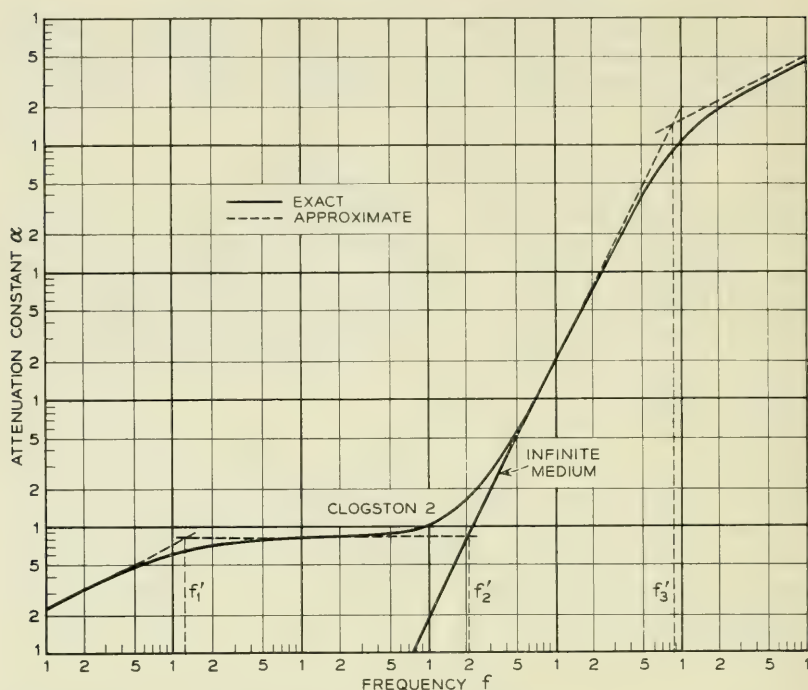


Fig. 20—Attenuation constants of a plane Clogston 2 line and an infinite laminated medium versus frequency on log-log scale.

numerical values are of no special significance, having been chosen solely for convenience in plotting.

So far as the practical applications of Clogston lines are concerned, we are primarily interested in the frequency range $f'_1 \leq f \leq f'_2$, where the attenuation constant is essentially independent of frequency. To determine the rate at which the attenuation constant of the p th mode begins to deviate from its "flat" value as the frequency is increased, we write equation (459) in the form

$$\begin{aligned} \alpha &= \frac{p^2 \pi^2}{2\sqrt{\bar{\mu}/\bar{\epsilon}} \bar{g} a^2} \left[1 + \frac{4t_1^2 T_1^2}{3p^2 \pi^2 \delta_1^4} \right] \\ &= \frac{p^2 \pi^2}{2\sqrt{\bar{\mu}/\bar{\epsilon}} \bar{g} a^2} \left[1 + \frac{4t_1^2 T_1^2 \mu_1^2 g_1^2 f^2}{3p^2} \right]. \end{aligned} \quad (468)$$

The two terms in the square brackets are equal, and the attenuation constant is double its "flat" value, when $f = f'_2$, a result which is in very good agreement with the calculated values shown in Fig. 20.

The maximum permissible thickness of the conducting layers in a

plane Clogston 2 with high-impedance walls, if the attenuation constant of the p th mode is not to exceed its "flat" value α_0 by more than a specified small fraction $\Delta\alpha/\alpha_0$ at a top frequency f_m , is easily shown from (468) to be

$$t_1 = \frac{\sqrt{3} p}{2T_1\mu_1g_1f_m} \sqrt{\frac{\Delta\alpha}{\alpha_0}}. \quad (469)$$

Measuring f_m in $\text{Mc} \cdot \text{sec}^{-1}$ and thicknesses in mils, and putting in numerical values for copper, we obtain

$$(t_1)_{\text{mils}} = \frac{36.84p}{(2T_1)_{\text{mils}}(f_m)_{\text{Mc}}} \sqrt{\frac{\Delta\alpha}{\alpha_0}}. \quad (470)$$

We see from (461) that for fixed θ , α_0 is inversely proportional to $(a/p)^2$, where a is the total thickness of the stack and p is the mode number, while from (469) the permissible value of t_1 for a certain fractional change in attenuation is inversely proportional to (a/p) , and also inversely proportional to the top frequency f_m .

It is interesting to compare equation (470) for the principal mode ($p = 1$) with equation (199) of Section V for the principal mode in an extreme Clogston 1 line with copper conducting layers. Since in a plane line $\Delta R/R_0 = \Delta\alpha/\alpha_0$, equation (199) may be written

$$(t_1)_{\text{mils}} = \frac{40.62}{(2T_1)_{\text{mils}}(f_m)_{\text{Mc}}} \sqrt{\frac{\Delta\alpha}{\alpha_0}}, \quad (471)$$

where $2T_1$ represents the total thickness of copper in both stacks. We expect that for partially filled plane Clogston lines with different proportions of the available space occupied by stacks, the maximum permissible layer thickness will be given by equations similar to (470) and (471), with values of the numerical coefficient intermediate between 36.84 and 40.62.

We turn next to a discussion of coaxial Clogston cables with finite laminae. A coaxial Clogston 2 is shown schematically in Fig. 11, and an enlarged view of part of the laminated stack in Fig. 4. The boundary conditions which apply to circular transverse magnetic waves on this structure are satisfied if at every point the sum of the radial impedances looking in opposite directions is zero. If we knew the explicit relation between the impedances at the two surfaces of the stack in terms of the stack parameters and the propagation constant γ , the impedance-matching conditions at the inner core and the outersheath would yield a transcendental equation for the propagation constants of

the various possible modes. If the coaxial layers are of finite thickness, however, the relation between the surface impedances of the stack involves the product of as many different matrices as there are layers in the whole stack, and this matrix product is not suited to analytic treatment. We shall therefore approach the problem from another point of view.

We have seen that if the conducting layers in a laminated transmission line are sufficiently thin compared to the skin depth, the attenuation constant is essentially independent of frequency. In practice it is important to know how rapidly the attenuation constant of a Clogston cable with finite laminae begins to deviate from its low-frequency value as the frequency is increased. In accordance with the results for the parallel-plane line, we expect the initial increase to be proportional to the square of the frequency. We shall derive the term proportional to f^2 in the attenuation constant of the coaxial line on the basis of the following assumptions:

We assume that the macroscopic current distribution in a coaxial Clogston 2 is independent of frequency, and hence is given by the expressions which have already been derived for the case of infinitesimally thin laminae. (It is easy to show that this assumption is valid for a *plane* Clogston 2.) If the conducting layers are of finite thickness, then each carries a definite finite fraction of the total current in the line. At low frequencies the current density in any given layer is approximately uniform, but as the frequency is increased it becomes nonuniform because of the development of skin effect, and the power dissipated in the layer is increased. We shall calculate the total power dissipated in the stack, and the corresponding attenuation constant, up to terms in f^2 .

Let the j th conducting layer in the stack be a hollow cylinder of conductivity g_1 , inner radius ρ_{j-1} , and thickness t_1 . Thus if there are $2n$ double layers we have $\rho_0 = a$ and $\rho_{2n} = b$, where as usual a and b denote the inner and outer radii of the whole stack. Let the total current flowing in the positive z -direction inside $\rho = \rho_{j-1}$ be I_{j-1} , and let the current flowing in the j th conducting layer be ΔI_j . It is shown in Appendix III that the average power dissipated per unit length in the j th conductor is approximately

$$\Delta P_j = \frac{1}{4\pi g_1 t_1 \rho_{j-1}} \left[|\Delta I_j|^2 + \frac{t_1^4}{3\delta_1^4} |I_{j-1}|^2 \right], \quad (472)$$

up to terms in $(t_1/\delta_1)^4$, where curvature corrections of the order of t_1/ρ_{j-1} have been neglected in comparison with unity. Presumably the only layers for which it may not be justifiable to neglect curvature cor-

reactions will be the extreme inner layers, which occupy at most a small fraction of the total volume of the stack.

The average current density \bar{J}_z in a Clogston 2 with infinitesimally thin laminae is given by equation (305) of Section VIII; namely, writing χ_p for the p th mode and dropping $e^{-\gamma z}$,

$$\bar{J}_z = H_0 \chi_p C'_0(\chi_p \rho), \quad (473)$$

where H_0 is an arbitrary amplitude constant. For $n = 0$ and 1, $C_n(\chi_p \rho)$ denotes the combination of Bessel functions

$$C_n(\chi_p \rho) = N_1(\chi_p b) J_n(\chi_p \rho) - J_1(\chi_p b) N_n(\chi_p \rho); \quad (474)$$

and χ_p is the p th positive root of

$$C_1(\chi a) = N_1(\chi b) J_1(\chi a) - J_1(\chi b) N_1(\chi a) = 0. \quad (475)$$

According to equation (434) of Section X, we may write

$$\chi_p = \frac{p\pi f_p(a/b)}{b-a}, \quad (476)$$

where the functions $f_p(a/b)$ are of the order of unity. The total current $I(\rho)$ flowing in the positive z -direction between the inner core and a cylinder of arbitrary radius ρ is just

$$I(\rho) = 2\pi \int_a^\rho \rho \bar{J}_z d\rho = 2\pi H_0 \rho C_1(\chi_p \rho). \quad (477)$$

The thickness of the j th conducting layer in a stack of finite layers may be written

$$t_1 = \theta(t_1 + t_2) = \theta(\rho_j - \rho_{j-1}) = \theta \Delta \rho_j, \quad (478)$$

where $\Delta \rho_j$ represents the thickness $t_1 + t_2$ of the j th double layer. Hence approximately

$$\Delta I_j = 2\pi \rho_{j-1} \bar{J}_z \Delta \rho_j = 2\pi H_0 \chi_p \rho_{j-1} C_0(\chi_p \rho_{j-1}) \Delta \rho_j, \quad (479)$$

it being remembered that the conduction current in the conducting layer is essentially equal to the total current in the double layer, since the displacement currents are negligible. The current flowing inside the radius ρ_{j-1} is, from (477),

$$I_{j-1} = 2\pi H_0 \rho_{j-1} C_1(\chi_p \rho_{j-1}), \quad (480)$$

and so the power dissipated per unit length in the j th conductor is

given by (472) as

$$\Delta P_j = \frac{\pi H_0 H_0^* \rho_{j-1}}{\theta g_1} \left[\chi_p^2 C_0^2(\chi_p \rho_{j-1}) + \frac{\theta^2 t_1^2}{3\delta_1^4} C_1^2(\chi_p \rho_{j-1}) \right] \Delta \rho_j. \quad (481)$$

The total power ΔP dissipated per unit length in the whole stack is obtained by summing ΔP_j over j . Approximately the sum by an integral, we have

$$\begin{aligned} \Delta P &= \frac{\pi H_0 H_0^*}{\bar{g}} \int_a^b \rho \left[\chi_p^2 C_0^2(\chi_p \rho) + \frac{\theta^2 t_1^2}{3\delta_1^4} C_1^2(\chi_p \rho) \right] d\rho \\ &= \frac{\pi H_0 H_0^* \chi_p^2}{2\bar{g}} \left[1 + \frac{\theta^2 t_1^2}{3\chi_p^2 \delta_1^4} \right] [b^2 C_0^2(\chi_p b) - a^2 C_0^2(\chi_p a)]. \end{aligned} \quad (482)$$

The average transmitted power P when the laminae are infinitesimally thin is

$$\begin{aligned} P &= \text{Re } \frac{1}{2} \int_0^{2\pi} \int_a^b \bar{E}_\rho H_{\phi\rho}^* d\rho d\phi \\ &= \pi \sqrt{\frac{\bar{\mu}}{\bar{\epsilon}}} H_0 H_0^* \int_a^b \rho C_1^2(\chi_p \rho) d\rho \\ &= \frac{1}{2} \pi \sqrt{\frac{\bar{\mu}}{\bar{\epsilon}}} H_0 H_0^* [b^2 C_0^2(\chi_p b) - a^2 C_0^2(\chi_p a)]. \end{aligned} \quad (483)$$

If we assume the same value for P when the laminae are of finite thickness, then from (482) and (483) the attenuation constant of the line is

$$\alpha = \frac{\Delta P}{2P} = \frac{\chi_p^2}{2\sqrt{\bar{\mu}/\bar{\epsilon}} \bar{g}} \left[1 + \frac{\theta^2 t_1^2}{3\chi_p^2 \delta_1^4} \right]. \quad (484)$$

The similarity of equation (484) to equation (468) for the parallel-plane line becomes obvious if we write χ_p in the form (476) and denote the total thickness $\theta(b-a)$ of conducting material in the coaxial stack by $2T_1$. We then have

$$\begin{aligned} \alpha &= \frac{p^2 \pi^2 f_p^2(a/b)}{2\sqrt{\bar{\mu}/\bar{\epsilon}} \bar{g} (b-a)^2} \left[1 + \frac{4t_1^2 T_1^2}{3p^2 \pi^2 f_p^2(a/b) \delta_1^4} \right] \\ &= \frac{p^2 \pi^2 f_p^2(a/b)}{2\sqrt{\bar{\mu}/\bar{\epsilon}} \bar{g} (b-a)^2} \left[1 + \frac{4t_1^2 T_1^2 \mu_1^2 g_1^2 f^2}{3p^2 f_p^2(a/b)} \right], \end{aligned} \quad (485)$$

and as the ratio a/b approaches unity the function $f_p^2(a/b)$ approaches unity and (485) becomes identical with (468). We recall that $f_1^2(a/b)$ was plotted against a/b in Fig. 12. For the principal mode in a cable

with no inner core ($a = 0$), equation (485) takes the form

$$\alpha = \frac{7.341}{\sqrt{\bar{\mu}_1 \epsilon} \bar{g} b^2} [1 + 0.8963 t_1^2 T_1^2 \mu_1^2 g_1^2 f^2]. \quad (486)$$

It should be emphasized that whereas equation (468) was obtained from a rigorous solution of the boundary-value problem for the plane line, equation (485) for the coaxial cable has been derived on the basis of certain physical assumptions and approximations whose effect on the accuracy of the final result is not very easy to estimate. Presumably one might check the accuracy of (485) for a particular Clogston cable by setting up the matrix relation between the known surface impedances of the core and the outer sheath and solving numerically for the propagation constant. It should not be too difficult to solve the matrix equation by cut-and-try methods for a cable having, say, two hundred double layers, if the matrix of each double layer were assumed to be given by equations (88) of Section III, and high-speed computing machinery were used to perform the matrix multiplications. In the absence of any such numerical results, however, we shall merely assume that equation (485) furnishes a reasonable approximation to the attenuation constant of a coaxial Clogston 2 in the frequency range $f'_1 \leq f \leq f'_3$, where f'_1 and f'_3 are the critical frequencies defined by (467).

The first conclusion which we can draw from (485) is that the maximum permissible thickness of the conducting layers in a coaxial Clogston 2 with high-impedance boundaries, if the attenuation constant of the p th mode is not to exceed its "flat" value α_0 by more than a specified small fraction $\Delta\alpha'/\alpha_0$ at a top frequency f_m , is

$$t_1 = \frac{\sqrt{3} p f_p(a/b)}{2 T_1 \mu_1 g_1 f_m} \sqrt{\frac{\Delta\alpha}{\alpha_0}}; \quad (487)$$

or, putting in numerical values for copper,

$$(t_1)_{\text{mils}} = \frac{36.84 p f_p(a/b)}{(2 T_1)_{\text{mils}} (f_m)_{\text{Mc}}} \sqrt{\frac{\Delta\alpha}{\alpha_0}}. \quad (488)$$

For the principal mode in a Clogston cable with no inner core, this becomes

$$(t_1)_{\text{mils}} = \frac{44.93}{(2 T_1)_{\text{mils}} (f_m)_{\text{Mc}}} \sqrt{\frac{\Delta\alpha}{\alpha_0}}. \quad (489)$$

As a second application of equation (485), we shall determine the upper crossover frequency at which the attenuation constant of a Clogston 2 is equal to the attenuation constant of a conventional coaxial

cable of the same size. Since the lower crossover frequency was found at the end of Section VIII, we shall then know the theoretical limits of the frequency range over which a given Clogston cable can have lower loss than the corresponding standard coaxial.

According to equation (317) of Section VIII, a conventional coaxial cable of radius b and optimum proportions has an attenuation constant

$$\alpha = \frac{1.796}{\sqrt{\mu_0/\epsilon_0} g_1 \delta_1 b} = \frac{1.796 \sqrt{\pi \mu_1 g_1 f}}{\sqrt{\mu_0/\epsilon_0} g_1 b}. \quad (490)$$

We shall assume that the upper crossover occurs in the high-frequency range where the attenuation constant of a Clogston 2 is approximately proportional to f^2 . Then for the p th mode in a cable with no inner core ($a = 0$), equation (485) gives

$$\alpha = \frac{2\pi^2 t_1^2 T_1^2 \mu_1^2 g_1^2 f^2}{3\sqrt{\bar{\mu}/\bar{\epsilon}} \bar{g} b^2} = \frac{\pi^2 t_1^2 \mu_1^2 \bar{g} f^2}{6\sqrt{\bar{\mu}/\bar{\epsilon}}}. \quad (491)$$

A little algebra shows that the two attenuation constants are equal when

$$f = \frac{1}{\pi \mu_1 g_1} \left[\frac{10.77}{2T_1 t_1^2} \sqrt{\frac{\bar{\mu} \epsilon_0}{\mu_0 \bar{\epsilon}}} \right]^{\frac{2}{3}}. \quad (492)$$

If the conventional cable is air-filled, then assuming copper conductors and no magnetic materials, we find that equation (492) becomes, numerically,

$$f_{Mc} = 33.02 \left[\frac{1}{(2T_1)_{\text{mils}} (t_1^2)_{\text{mils}}} \sqrt{\frac{1-\theta}{\epsilon_{2r}}} \right]^{\frac{2}{3}}. \quad (493)$$

If we consider a 3/8-inch Clogston cable with 0.1-mil copper conductors, 0.05-mil polyethylene insulators, and no inner core, then

$$\begin{aligned} b &= 187.5 \text{ mils} & \ell &= 2/3 \\ 2T_1 &= 125 \text{ mils} & \epsilon_{2r} &= 2.26 \\ t_1 &= 0.1 \text{ mils} \end{aligned} \quad (494)$$

We found in Section VIII that the lower crossover frequency for this cable is about $50 \text{ kc} \cdot \text{sec}^{-1}$, while from equation (493) the upper crossover frequency turns out to be $15 \text{ Mc} \cdot \text{sec}^{-1}$.

We next discuss the problem of maximizing the frequency band over which the attenuation constant of a Clogston cable of given diameter

does not exceed a specified value.²⁶ We suppose that the thickness t_1 of the conductors is fixed, but that the proportion of conducting material in the cable may be adjusted by changing the thickness of the insulators. Let α_m be the value of the attenuation constant which is not to be exceeded in the operating frequency range, and let f_m be the frequency at which this maximum attenuation is reached. What should be the fraction θ of conducting material in the cable in order to maximize f_m ? It is tacitly assumed that α_m is at least slightly greater than the minimum "flat" attenuation constant which is possible with a cable of the given diameter, since obviously we do not wish to work entirely in the very-low-frequency range.

In the frequency range of interest the attenuation constant of the p th mode is given by equation (484), which may be written

$$\alpha = \frac{\chi_p^2}{2\sqrt{\bar{\mu}/\bar{\epsilon}}\bar{g}} + \frac{\theta^2 t_1^2 \pi^2 \mu_1^2 g_1^2 f^2}{6\sqrt{\bar{\mu}/\bar{\epsilon}}\bar{g}}, \quad (495)$$

where χ_p is a root of (475) and independent of θ . Solving (495) for the frequency f_m at which α is equal to α_m , and substituting for $\bar{\epsilon}$, $\bar{\mu}$, and \bar{g} from (268), we obtain

$$f_m = \frac{\sqrt{3}}{\pi \mu_1 g_1 t_1} \left[\frac{2[\theta \mu_1 + (1 - \theta) \mu_2]^{\frac{1}{2}} [(1 - \theta)/\epsilon_2]^{\frac{1}{2}} g_1 \alpha_m - \chi_p^2}{\theta} \right]^{\frac{1}{2}}. \quad (496)$$

A routine calculation shows that f_m is a maximum, considered as a function of θ , when θ satisfies

$$\frac{\alpha_m g_1 \theta [\theta \mu_1 + 2(1 - \theta) \mu_2]}{[\theta \mu_1 + (1 - \theta) \mu_2]^{\frac{1}{2}} (1 - \theta)^{\frac{1}{2}} \epsilon_2^{\frac{1}{2}}} = 2\chi_p^2. \quad (497)$$

Equation (497) is easily reduced to a quartic equation in θ , which may be solved without difficulty when the other parameters are given. The maximum value of f_m is then obtained by substituting θ back into (496).

We shall now investigate in more detail the case in which

$$\mu_1 = \mu_2, \quad (498)$$

that is, the permeabilities of the conducting and insulating layers are equal. In this case the low-frequency attenuation constant α_0 , which is just the first term on the right side of equation (495), is given by

$$\alpha_0 = \frac{\chi_p^2}{2\theta \sqrt{1 - \theta} \sqrt{\mu_2} \epsilon_2 g_1}, \quad (499)$$

²⁶ A similar problem was first investigated in an unpublished memorandum by H. S. Black.

and α_0 has a minimum when

$$\theta = 2/3. \quad (500)$$

The minimum value of the low-frequency attenuation constant, which we may call α_{00} , is just

$$\alpha_{00} = \frac{3\sqrt{3} \chi_p^2}{4\sqrt{\mu_2/\epsilon_2} g_1}. \quad (501)$$

Writing

$$\chi_p^2 = \frac{4\sqrt{\mu_2/\epsilon_2} g_1 \alpha_{00}}{3\sqrt{3}}, \quad (502)$$

we find that equation (496) takes the form

$$f_m = \frac{\sqrt{3} \chi_p}{\pi \mu_1 g_1 t_1} \left[\frac{3\sqrt{3} (1 - \theta)^{\frac{1}{2}}}{2\theta} \frac{\alpha_m}{\alpha_{00}} - \frac{1}{\theta^2} \right]^{\frac{1}{2}}, \quad (503)$$

for any value of θ . From equation (497), f_m is a maximum when θ satisfies

$$\frac{\theta(2 - \theta)}{(1 - \theta)^{\frac{3}{2}}} = \frac{8}{3\sqrt{3}} \frac{\alpha_{00}}{\alpha_m}, \quad (504)$$

which is equivalent to the quartic equation

$$\theta^4 - 4\theta^3 + 4\theta^2 + \frac{64\alpha_{00}^2}{27\alpha_m^2} (\theta - 1) = 0. \quad (505)$$

If θ_m is the root of (505) which lies between zero and one, then the corresponding value of f_m is

$$f_m = \frac{\sqrt{3} \chi_p}{\pi \mu_1 g_1 t_1 \theta_m} \left[\frac{2 - 3\theta_m}{2 - \theta_m} \right]^{\frac{1}{2}}. \quad (506)$$

We observe from either (503) or (506) that f_m is inversely proportional to t_1 .

The values of θ_m and $\theta_m^{-1}[(2 - 3\theta_m)/(2 - \theta_m)]^{\frac{1}{2}}$ are plotted in Fig. 21 against α_m/α_{00} , which is just the ratio of the maximum attenuation constant to the minimum low-frequency attenuation constant which can be achieved with a Clogston cable of the same diameter. When α_m/α_{00} is unity, then $\theta_m = 2/3$ and f_m is zero to the present approximation (a better estimate of f_m would be the critical frequency f_1' defined by equation (467)). For values of α_m/α_{00} greater than about five, θ_m is

given to a good approximation by

$$\theta_m \approx \frac{4\alpha_{00}}{3\sqrt{3}\alpha_m} \approx \frac{0.77\alpha_{00}}{\alpha_m}, \quad (507)$$

while

$$f_m \approx \frac{2.25\chi_p}{\pi\mu_1 g_1 l_1} \frac{\alpha_m}{\alpha_{00}}. \quad (508)$$

The low-frequency attenuation constant α_0 of a Clogston cable with $\theta = \theta_m$ will of course be greater than α_{00} if θ_m is not equal to $2/3$. This is not really a disadvantage, however, since by assumption we only wish to insure that $\alpha \leq \alpha_m$ over the operating band, and the nearer α approaches to α_m over the whole band the less serious will be the equalization problem. It may be shown that the ratio α_0/α_m decreases from unity toward one-half as α_m/α_{00} is increased indefinitely. Physically this means that the low-frequency attenuation constant of an optimum Clogston cable is always at least half as great as the attenuation constant at the upper end of the band, and the cable never contains more conducting material than would correspond to a total stack thickness of about two effective skin depths at the highest operating frequency.

We conclude with a few numerical formulas relating to the principal mode in a completely filled Clogston cable with copper conductors and no inner core. The low-frequency attenuation constant α_0 of such a

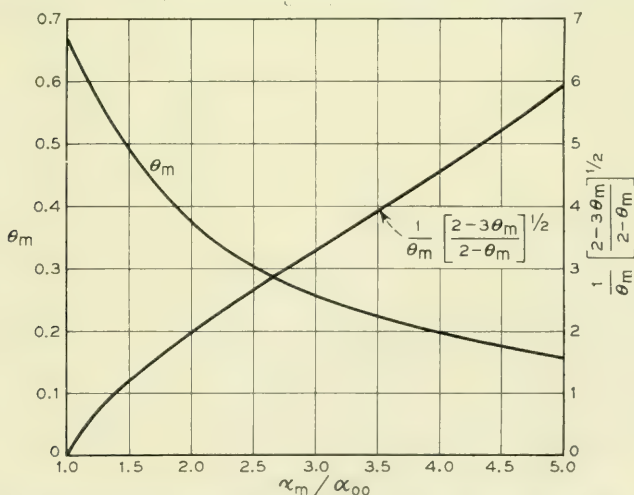


Fig. 21—Curves related to the optimum fraction θ_m of conducting material in Clogston cables with finite laminae, as a function of the attenuation ratio α_m/α_{00} .

cable is given by

$$\begin{aligned}\alpha_0 &= \frac{0.521\sqrt{\epsilon_{2r}}}{\theta\sqrt{1-\theta}b_{\text{mils}}^2} \text{ nepers} \cdot \text{meter}^{-1} \\ &= \frac{0.728 \times 10^4 \sqrt{\epsilon_{2r}}}{\theta\sqrt{1-\theta}b_{\text{mils}}^2} \text{ db} \cdot \text{mile}^{-1},\end{aligned}\quad (509)$$

for any value of θ , while if $\theta = 2/3$, we have

$$\begin{aligned}\alpha_{00} &= \frac{1.353\sqrt{\epsilon_{2r}}}{b_{\text{mils}}^2} \text{ nepers} \cdot \text{meter}^{-1} \\ &= \frac{1.891 \times 10^4 \sqrt{\epsilon_{2r}}}{b_{\text{mils}}^2} \text{ db} \cdot \text{mile}^{-1}.\end{aligned}\quad (510)$$

The frequency f_m as a function of the ratio α_m/α_{00} is

$$(f_m)_{\text{Mc}} = \frac{44.93}{(t_1)_{\text{mils}}b_{\text{mils}}} \left[2.598 \frac{(1-\theta)^{\frac{1}{2}}}{\theta} \frac{\alpha_m}{\alpha_{00}} - \frac{1}{\theta^2} \right]^{\frac{1}{2}}, \quad (511)$$

for any θ , and when $\theta = \theta_m$ the expression for f_m becomes

$$(f_m)_{\text{Mc}} = \frac{44.93}{(t_1)_{\text{mils}}b_{\text{mils}}} \frac{1}{\theta_m} \left[\frac{2-3\theta_m}{2-\theta_m} \right]^{\frac{1}{2}}, \quad (512)$$

where the factor involving θ_m is plotted against α_m/α_{00} in Fig. 21.

If we consider a 3/8-inch Clogston cable with 0.1-mil copper conductors and polyethylene insulation ($\epsilon_{2r} = 2.26$), we find

$$\alpha_{00} = 0.809 \text{ db} \cdot \text{mile}^{-1}. \quad (513)$$

If we set

$$\alpha_m = 2\alpha_{00} = 1.618 \text{ db} \cdot \text{mile}^{-1}, \quad (514)$$

then it turns out that

$$\theta_m = 0.3745, \quad (515)$$

so that the insulating layers should be 0.167 mil thick. The low-frequency attenuation constant for θ_m is

$$\alpha_0 = 1.300\alpha_{00} = 1.051 \text{ db} \cdot \text{mile}^{-1}, \quad (516)$$

and α_m is reached at a frequency

$$f_m = 4.70 \text{ Mc} \cdot \text{sec}^{-1}. \quad (517)$$

If we had used $\theta = 2/3$, we should have reached α_m at a frequency of

$3.59 \text{ Me} \cdot \text{sec}^{-1}$, which is only 76.5 per cent of the frequency given by (517). An ordinary air-filled coaxial cable of the same size would have an attenuation constant equal to α_{00} at about $50 \text{ ke} \cdot \text{sec}^{-1}$ and equal to $\alpha_m (= 2\alpha_{00})$ at about $200 \text{ ke} \cdot \text{sec}^{-1}$.

It should be borne in mind that in the preceding example we have neglected the effects of dielectric loss and of stack nonuniformity. Neither of these effects can be completely eliminated in a physical Clogston cable, and both will exert increasingly adverse influences on the attenuation constant as the frequency is raised.

XII. EFFECT OF NONUNIFORMITY OF LAMINATED MEDIUM

In the previous analysis of laminated transmission lines we have treated only perfectly uniform structures, in which every conducting layer is identical to every other conducting layer in thickness and in electrical properties, and all the insulating layers are similarly identical to each other. In practice, however, it will not be possible to lay down large numbers of absolutely identical thin layers, and we therefore need to know the effect on transmission of slight nonuniformities in the laminated stacks. Some indication that stack uniformity will be a very critical problem in laminated cables which are expected to give large improvements in attenuation over conventional coaxial cables of the same size may be obtained from the results of Section VI, which showed that in a Clogston 1 line, where the phase velocity is determined by the $\mu\epsilon$ product of the main dielectric, this product must be controlled very accurately to maintain the desired deep penetration of current into the laminated stacks. In a Clogston 2, where the main dielectric has been replaced by extensions of the stacks, one might expect similarly stringent requirements on the uniformity of the laminated material if the desired current distribution is to be maintained.

In this section we estimate the effects of stack nonuniformity by studying some particular idealized cases of nonuniformity in a parallel-plane Clogston 2 with infinitesimally thin layers, in which the average electrical properties of the stack vary only in the direction perpendicular to the layers. The principal conclusion is that if one attempts to realize with a Clogston line an attenuation constant which is a small fraction, say of the order of one-tenth, of the attenuation constant of a conventional line of the same dimensions at the same frequency, then long-range variations in the properties of the stack (as distinguished from short-range random fluctuations) must be controlled to within a few parts in 10,000. The price is less steep if the overall improvement sought

is less, but in all practical cases it appears that the average properties of the stack must be held constant against slow variations to a fraction of a per cent. The requirement of extraordinarily high precision is in addition to the requirement that the individual layers must be extremely thin if a Clogston cable is to improve on a conventional coaxial cable at all in the megacycle frequency range.

For purposes of analysis, we consider a parallel-plane Clogston 2 transmission line bounded by infinite-impedance sheaths at $y = \pm \frac{1}{2}a$, as shown schematically in Fig. 10. The individual layers are supposed to be infinitesimally thin, so that near any given point the average electrical constants of the stack are

$$\begin{aligned}\bar{\epsilon} &= \epsilon_2/(1 - \theta), \\ \bar{\mu} &= \theta\mu_1 + (1 - \theta)\mu_2, \\ \bar{g} &= \theta g_1.\end{aligned}\tag{518}$$

The quantities $\bar{\epsilon}$, $\bar{\mu}$, and \bar{g} may vary, continuously or with a finite number of finite discontinuities, as functions of the transverse coordinate y , owing to variations in any or all of μ_1 , g_1 , μ_2 , ϵ_2 , and θ ; but they are not supposed to vary with x or z .

We shall be concerned with modes in which the fields are independent of x , and in which the only field components are H_x , \bar{E}_y , and E_z . Then Maxwell's equations are given by (269) of Section VIII, and reduce, if we write the field components in the form $H_x(y)e^{-\gamma z}$, $\bar{E}_y(y)e^{-\gamma z}$, and $E_z(y)e^{-\gamma z}$, to

$$\begin{aligned}-\gamma H_x &= i\omega\bar{\epsilon}\bar{E}_y, \\ dH_x/dy &= -\bar{g}E_z, \\ -\gamma\bar{E}_y - dE_z/dy &= i\omega\bar{\mu}H_x.\end{aligned}\tag{519}$$

If we eliminate \bar{E}_y and E_z from these equations we obtain

$$\frac{d^2 H_x}{dy^2} - \frac{1}{\bar{g}} \frac{d\bar{g}}{dy} \frac{dH_x}{dy} - i\omega\bar{\mu}\bar{g} \left[1 + \frac{\gamma^2}{\omega^2\bar{\mu}\bar{\epsilon}} \right] H_x = 0,\tag{520}$$

where H_x and E_z must be continuous at any points of discontinuity of $\bar{\epsilon}$, $\bar{\mu}$, or \bar{g} . The tangential magnetic field must vanish on the infinite-impedance surfaces at $y = \pm \frac{1}{2}a$; hence we have the boundary conditions

$$H_x(-\tfrac{1}{2}a) = H_x(\tfrac{1}{2}a) = 0.\tag{521}$$

These boundary conditions, taken in conjunction with the differential

equation (520), define the values of γ which are the propagation constants of the various modes of the line.

While it is possible to find special forms of the functions $\bar{\epsilon}(y)$, $\bar{\mu}(y)$, and $\bar{g}(y)$ such that (520) can be solved exactly in terms of known functions, it is easier to make certain approximations in the beginning which retain only the important terms. For this purpose we shall write

$$\begin{aligned}\bar{\epsilon} &= \bar{\epsilon}_0 + \Delta\bar{\epsilon}, \\ \bar{\mu} &= \bar{\mu}_0 + \Delta\bar{\mu}, \\ \bar{g} &= \bar{g}_0 + \Delta\bar{g},\end{aligned}\tag{522}$$

where $\bar{\epsilon}_0$, $\bar{\mu}_0$, and \bar{g}_0 are constants representing the average values of $\bar{\epsilon}$, $\bar{\mu}$, and \bar{g} across the stack, so that the average values of $\Delta\bar{\epsilon}$, $\Delta\bar{\mu}$, and $\Delta\bar{g}$ across the stack are zero.* Furthermore the fractional variations in the stack parameters will be assumed small compared to unity; in practical cases they will never be larger than a few per cent and will usually be only a fraction of one per cent.

Referring now to equation (520), we see that the coefficient of H_x contains the large factor $\omega\bar{\mu}\bar{g}$, which is of the order of $1/\delta_1^2$, as compared with the term d^2H_x/dy^2 , which is presumably of the order of $(1/a^2)H_x$. Hence small changes in $\bar{\epsilon}$ and $\bar{\mu}$ will make relatively large changes in the coefficient of H_x , since γ^2 is a constant. On the other hand, the coefficient of dH_x/dy will be small for any reasonable variations in the small quantity $\Delta\bar{g}/\bar{g}_0$. Hence we shall neglect this term entirely and deal with the approximate equation

$$\frac{d^2H_x}{dy^2} - i\omega\bar{\mu}\bar{g} \left[1 + \frac{\gamma^2}{\omega^2\bar{\mu}\bar{\epsilon}} \right] H_x = 0.\tag{523}$$

If we substitute (522) into (523) and drop second order terms in $\Delta\bar{\epsilon}/\bar{\epsilon}_0$, $\Delta\bar{\mu}/\bar{\mu}_0$, and $\Delta\bar{g}/\bar{g}_0$, we find that the coefficient of H_x becomes

$$\begin{aligned}& -\frac{i\bar{g}}{\omega\bar{\epsilon}} [\omega^2\bar{\mu}\bar{\epsilon} + \gamma^2] \\ & \approx -\frac{i\bar{g}_0}{\omega\bar{\epsilon}_0} \left[1 + \frac{\Delta\bar{g}}{\bar{g}_0} - \frac{\Delta\bar{\epsilon}}{\bar{\epsilon}_0} \right] \left[\gamma^2 + \omega^2\bar{\mu}_0\bar{\epsilon}_0 + \omega^2\bar{\mu}_0\bar{\epsilon}_0 \left(\frac{\Delta\bar{\mu}}{\bar{\mu}_0} + \frac{\Delta\bar{\epsilon}}{\bar{\epsilon}_0} \right) \right];\end{aligned}\tag{524}$$

and if

$$\Gamma_t^2 = \frac{i\bar{g}_0}{\omega\bar{\epsilon}_0} [\omega^2\bar{\mu}_0\bar{\epsilon}_0 + \gamma^2],\tag{525}$$

* The present use of zero subscripts on $\bar{\epsilon}_0$, $\bar{\mu}_0$, and \bar{g}_0 has of course nothing to do with the earlier convention that associated zero subscripts with the main dielectric in Clogston 1 lines.

so that

$$\gamma^2 = -\omega^2 \bar{\mu}_0 \bar{\epsilon}_0 - (i\omega \bar{\epsilon}_0 / \bar{g}_0) \Gamma_t^2, \quad (526)$$

then (524) becomes, approximately,

$$\begin{aligned} & - \left[1 + \frac{\Delta \bar{g}}{\bar{g}_0} - \frac{\Delta \bar{\epsilon}}{\bar{\epsilon}_0} \right] \left[\Gamma_t^2 + i\omega \bar{\mu}_0 \bar{g}_0 \left(\frac{\Delta \bar{\mu}}{\bar{\mu}_0} + \frac{\Delta \bar{\epsilon}}{\bar{\epsilon}_0} \right) \right] \\ & \approx - \left[\Gamma_t^2 + i\omega \bar{\mu}_0 \bar{g}_0 \left(\frac{\Delta \bar{\mu}}{\bar{\mu}_0} + \frac{\Delta \bar{\epsilon}}{\bar{\epsilon}_0} \right) \right]. \end{aligned} \quad (527)$$

In all cases of interest we shall find that $(\Delta \bar{\mu} / \bar{\mu}_0 + \Delta \bar{\epsilon} / \bar{\epsilon}_0)$ is smaller than or at most of the same order of magnitude as $\Gamma_t^2 / i\omega \bar{\mu}_0 \bar{g}_0$. Hence the differential equation (523) takes the approximate form

$$\frac{d^2 H_x}{dy^2} - \left[\Gamma_t^2 + i\omega \bar{\mu}_0 \bar{g}_0 \left(\frac{\Delta \bar{\mu}}{\bar{\mu}_0} + \frac{\Delta \bar{\epsilon}}{\bar{\epsilon}_0} \right) \right] H_x = 0, \quad (528)$$

where Γ_t^2 is determined by the two-point boundary conditions (521).

The variations of the stack parameters appear in (528) only in the term $(\Delta \bar{\mu} / \bar{\mu}_0 + \Delta \bar{\epsilon} / \bar{\epsilon}_0)$, which is some as yet unspecified function of y . For convenience we shall write this term in the form

$$\frac{\Delta \bar{\mu}}{\bar{\mu}_0} + \frac{\Delta \bar{\epsilon}}{\bar{\epsilon}_0} = \frac{C}{\omega \bar{\mu}_0 \bar{g}_0 a^2} \varphi(y), \quad (529)$$

where C is a dimensionless parameter and $\varphi(y)$ is a function whose average value over the stack is zero, and whose maximum absolute value will usually be of the order of unity. It is worth noting that if the conducting and insulating layers all have equal permeabilities, then (529) becomes

$$\frac{\Delta \bar{\epsilon}}{\bar{\epsilon}_0} = \frac{C \delta_1^2}{2a^2 \theta_0} \varphi(y), \quad (530)$$

where δ_1 is the skin depth in the average conducting layer and θ_0 is the average fraction of space filled with conducting material. If we solve the differential equation for different values of the scale factor C but the same $\varphi(y)$, we can calculate the effect of stack nonuniformities of the same type but different amplitudes, or the effect of nonuniformity in the same stack at different frequencies. In the latter case C is directly proportional to the frequency.

The final step in the transformation of the differential equation (528) will be to reduce it to dimensionless form by the substitutions

$$\begin{aligned}
 \xi &= y/a + \tfrac{1}{2}, \\
 w(\xi) &= H_x(y), \\
 f(\xi) &= \varphi(y), \\
 \Lambda &= -\Gamma_t^2 a^2.
 \end{aligned}
 \tag{531}$$

Then on making use of (529) we get

$$d^2 w/d\xi^2 + [\Lambda - iCf(\xi)]w(\xi) = 0, \tag{532}$$

with the boundary conditions

$$w(0) = w(1) = 0. \tag{533}$$

Once Λ has been determined for a particular mode, the propagation constant γ is obtained from (526) and (531), namely

$$\gamma = i\omega\sqrt{\bar{\mu}_0\bar{\epsilon}_0} [1 + \Lambda/i\omega\bar{\mu}_0\bar{g}_0a^2]^{\frac{1}{2}}. \tag{534}$$

Assuming as usual that the attenuation per radian is small, we find that the attenuation and phase constants are given by

$$\alpha = \text{Re } \gamma = \text{Re } \frac{\Lambda}{2\sqrt{\bar{\mu}_0/\bar{\epsilon}_0} \bar{g}_0a^2}, \tag{535}$$

$$\beta = \text{Im } \gamma = \omega\sqrt{\bar{\mu}_0\bar{\epsilon}_0} + \text{Im } \frac{\Lambda}{2\sqrt{\bar{\mu}_0/\bar{\epsilon}_0} \bar{g}_0a^2}. \tag{536}$$

The eigenvalues Λ of the differential equation (532) with boundary conditions (533) may be found analytically for some simple forms of $f(\xi)$, or numerically using a differential analyzer for any given $f(\xi)$ which does not fluctuate too rapidly. When $C = 0$, as in the case of a perfectly uniform stack, the eigenvalues are obviously

$$\Lambda_1 = \pi^2, \quad \Lambda_2 = 4\pi^2, \dots, \tag{537}$$

corresponding to the eigenfunctions

$$w_1 = \sin \pi\xi, \quad w_2 = \sin 2\pi\xi, \dots \tag{538}$$

As C varies continuously, we expect the eigenvalues and eigenfunctions to vary continuously in a manner depending on $f(\xi)$. In the following paragraphs we shall discuss the behavior of Λ_1 , and sometimes also Λ_2 , as a function of C for various simple types of nonuniformity.

(i) $f(\xi)$ constant except at single discontinuity. Let

$$f(\xi) = \begin{cases} -\frac{1}{2\xi_0}, & 0 \leq \xi < \xi_0, \\ \frac{1}{2(1-\xi_0)}, & \xi_0 < \xi \leq 1, \end{cases} \quad (539)$$

where ξ_0 is some fixed number between 0 and 1 but not, in the cases of interest, extremely close to either 0 or 1. Solutions of (532) satisfying the boundary conditions (533) are obviously

$$\begin{aligned} w(\xi) &= A \sin [\Lambda + iC/2\xi_0]^{\frac{1}{2}}\xi, & 0 \leq \xi < \xi_0, \\ w(\xi) &= B \sin [\Lambda - iC/2(1-\xi_0)]^{\frac{1}{2}}(1-\xi), & \xi_0 < \xi \leq 1, \end{aligned} \quad (540)$$

where A and B are arbitrary constants. The requirements that w and $dw/d\xi$ be continuous* at $\xi = \xi_0$ lead to the equations

$$\begin{aligned} A \sin [\Lambda + iC/2\xi_0]^{\frac{1}{2}}\xi_0 &= B \sin [\Lambda - iC/2(1-\xi_0)]^{\frac{1}{2}}(1-\xi_0), \\ A[\Lambda + iC/2\xi_0]^{\frac{1}{2}} \cos [\Lambda + iC/2\xi_0]^{\frac{1}{2}}\xi_0 &= -B[\Lambda - iC/2(1-\xi_0)]^{\frac{1}{2}} \cos [\Lambda - iC/2(1-\xi_0)]^{\frac{1}{2}}(1-\xi_0), \end{aligned} \quad (541)$$

which will be consistent if this characteristic equation is satisfied:

$$\frac{\tan [\Lambda + iC/2\xi_0]^{\frac{1}{2}}\xi_0}{[\Lambda + iC/2\xi_0]^{\frac{1}{2}}} + \frac{\tan [\Lambda - iC/2(1-\xi_0)]^{\frac{1}{2}}(1-\xi_0)}{[\Lambda - iC/2(1-\xi_0)]^{\frac{1}{2}}} = 0. \quad (542)$$

The roots in Λ of equation (542) are the eigenvalues of the problem; the eigenfunction corresponding to any given eigenvalue is given by equations (540) after the ratio B/A is determined from either of equations (541).

It is easy to show that when $C = 0$, the roots of (542) are $\Lambda_1 = \pi^2$, $\Lambda_2 = 4\pi^2$, \dots . For large values of C , representing relatively great differences between the two parts of the stack, physical considerations lead us to expect that there will be pairs of modes, one member of each pair being essentially confined to each part of the stack and having a propagation constant determined approximately by the width of that part. It may in fact be shown that the asymptotic expression for the eigenvalue of the mode which is essentially confined to the region $0 \leq \xi < \xi_0$ is

$$\Lambda \approx \frac{\pi^2}{\xi_0^2} \left[1 - 2\sqrt{\frac{(1-\xi_0)}{\xi_0 C}} \right] - i \left[\frac{C}{2\xi_0} - \frac{2\pi^2}{\xi_0^2} \sqrt{\frac{(1-\xi_0)}{\xi_0 C}} \right]; \quad (543)$$

* The continuity of $dw/d\xi$ is a consequence of the continuity of E_z , provided that we neglect any discontinuity in \bar{g} at $\xi = \xi_0$.

and the asymptotic expression for the eigenvalue of the mode which is essentially confined to $\xi_0 < \xi \leq 1$ is

$$\Lambda \approx \frac{\pi^2}{(1 - \xi_0)^2} \left[1 - 2\sqrt{\frac{\xi_0}{(1 - \xi_0)C}} \right] + i \left[\frac{C}{2(1 - \xi_0)} - \frac{2\pi^2}{(1 - \xi_0)^2} \sqrt{\frac{\xi_0}{(1 - \xi_0)C}} \right]. \quad (544)$$

It is clear that if $\xi_0 < \frac{1}{2}$ the latter mode has the smaller attenuation constant, while if $\xi_0 > \frac{1}{2}$ the former mode has the smaller attenuation constant.

It is not difficult, although the details will be omitted here, to investigate the behavior of the eigenvalues for small C and to show that no matter whether $\xi_0 < \frac{1}{2}$ or $\xi_0 > \frac{1}{2}$, the eigenvalue which starts from π^2 at $C = 0$ tends to the asymptotic value which has the smaller real part, so that this eigenvalue, whether its asymptotic form be given by (543) or (544), may be called Λ_1 . It appears that if $\xi_0 < \frac{1}{2}$, then $\text{Im } \Lambda_1$ is positive for positive C , while if $\xi_0 > \frac{1}{2}$, then $\text{Im } \Lambda_1$ is negative for positive C .

An interesting mathematical phenomenon appears when $\xi_0 = \frac{1}{2}$, so that the discontinuity in $f(\xi)$ is exactly at the center of the stack. In this case, when C is small Λ_1 and Λ_2 are both real, Λ_1 being somewhat greater than π^2 and Λ_2 somewhat less than $4\pi^2$. For a certain value of C the two eigenvalues coincide; this value is approximately

$$C = 17.9, \quad \Lambda_1 = \Lambda_2 = 25.6. \quad (545)$$

For larger values of C , Λ_1 and Λ_2 are complex conjugates (it seems to be immaterial which is which) whose asymptotic forms are given by (543) and (544) with $\xi_0 = \frac{1}{2}$.

Approximate values of Λ_1 and Λ_2 were found for the symmetric case, $\xi_0 = 0.5$, and for one unsymmetric case, $\xi_0 = 0.6$, on the Laboratories' general purpose analog computer for $0 \leq C \leq 100$, and were refined afterward by desk computation, using a method of successive approximations to solve equation (542). The real and imaginary parts of Λ_1/π^2 and Λ_2/π^2 are plotted in Fig. 22 for the symmetric case, where it should be noted that different vertical scales are used for $\text{Re } \Lambda/\pi^2$ and $\text{Im } \Lambda/\pi^2$. The corresponding eigenfunctions $w_1(\xi)$ and $w_2(\xi)$ are shown in Fig. 23 for $C = 0$, $C = 17.9$, which corresponds to equal eigenvalues, and $C = 100$. It will be recalled that $w(\xi)$ is equal to $H_x(y)$, and the other field components can be derived from H_x by equations (519) if desired. Fig. 24 shows plots of Λ_1/π^2 and Λ_2/π^2 for the unsymmetric case $\xi_0 = 0.6$.

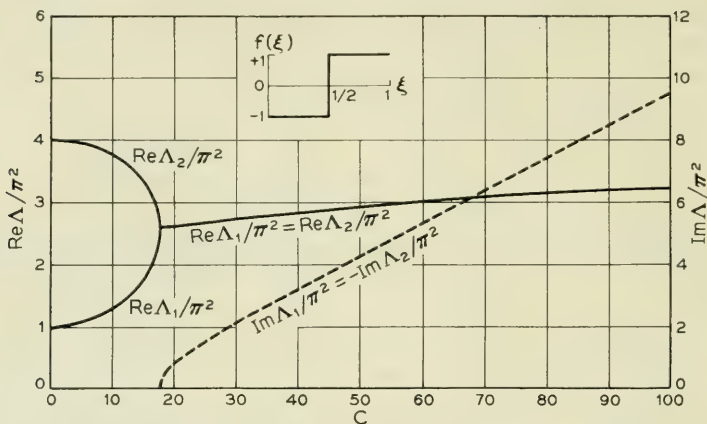


Fig. 22—Real and imaginary parts of Λ_1/π^2 and Λ_2/π^2 for a nonuniform stack whose average properties are constant except at a single, symmetric discontinuity.

(ii) $f(\xi)$ a symmetric rectangular step. Let

$$f(\xi) = \begin{cases} -\frac{1}{2\xi_0}, & 0 \leq \xi < \frac{1}{2}\xi_0, \\ \frac{1}{2(1 - \xi_0)}, & \frac{1}{2}\xi_0 < \xi < 1 - \frac{1}{2}\xi_0, \\ -\frac{1}{2\xi_0}, & 1 - \frac{1}{2}\xi_0 < \xi \leq 1, \end{cases} \quad (546)$$

where ξ_0 is some fixed number between 0 and 1 but not, in the cases of interest, extremely close to either 0 or 1. Inasmuch as $f(\xi)$ has even symmetry about $\xi = \frac{1}{2}$, every mode will preserve the (even or odd) symmetry about $\xi = \frac{1}{2}$ which it has when $C = 0$. We shall consider the lowest even mode,* which has the eigenfunction $\sin \pi\xi$ when $C = 0$. Solutions of (532) having even symmetry about $\xi = \frac{1}{2}$ (we need consider only the region $0 \leq \xi \leq \frac{1}{2}$ on account of the symmetry) and satisfying the boundary conditions (533) are given by

* For large C the lowest even mode will be confined essentially to

$$\frac{1}{2}\xi_0 < \xi < 1 - \frac{1}{2}\xi_0,$$

while the lowest odd mode will be confined to the two regions

$$0 \leq \xi < \frac{1}{2}\xi_0 \text{ and } 1 - \frac{1}{2}\xi_0 < \xi \leq 1.$$

If $\xi_0 > 2/3$, the latter mode will ultimately have a lower attenuation constant than the former; but we shall not take space to investigate it here.

$$\begin{aligned} w(\xi) &= A \sin [\Lambda + iC/2\xi_0]^{\frac{1}{2}}\xi, & 0 \leq \xi < \frac{1}{2}\xi_0, \\ w(\xi) &= B \cos [\Lambda - iC/2(1 - \xi_0)]^{\frac{1}{2}}(\frac{1}{2} - \xi), & \frac{1}{2}\xi_0 < \xi \leq \frac{1}{2}. \end{aligned} \quad (547)$$

The requirements that w and $dw/d\xi$ must be continuous at $\xi = \frac{1}{2}\xi_0$ lead to the equations

$$\begin{aligned} A \sin \frac{1}{2}[\Lambda + iC/2\xi_0]^{\frac{1}{2}}\xi_0 &= B \cos \frac{1}{2}[\Lambda - iC/2(1 - \xi_0)]^{\frac{1}{2}}(1 - \xi_0), \\ A[\Lambda + iC/2\xi_0]^{\frac{1}{2}} \cos \frac{1}{2}[\Lambda + iC/2\xi_0]^{\frac{1}{2}}\xi_0 &= B[\Lambda - iC/2(1 - \xi_0)]^{\frac{1}{2}} \sin \frac{1}{2}[\Lambda - iC/2(1 - \xi_0)]^{\frac{1}{2}}(1 - \xi_0), \end{aligned} \quad (548)$$

which will be consistent if the following characteristic equation is satisfied:

$$\frac{\tan \frac{1}{2}[\Lambda + iC/2\xi_0]^{\frac{1}{2}}\xi_0}{[\Lambda + iC/2\xi_0]^{\frac{1}{2}}} = \frac{\cot \frac{1}{2}[\Lambda - iC/2(1 - \xi_0)]^{\frac{1}{2}}(1 - \xi_0)}{[\Lambda - iC/2(1 - \xi_0)]^{\frac{1}{2}}}. \quad (549)$$

The roots in Λ of equation (549) are the eigenvalues corresponding to the even modes of the symmetrical structure.

When $C = 0$, the roots of (549) are $\Lambda = \pi^2, 9\pi^2, \dots$. It appears that for $C > 0$ we have $\text{Re } \Lambda_1 > \pi^2$ and $\text{Im } \Lambda_1 > 0$. For large C the asymptotic expression for Λ_1 turns out to be

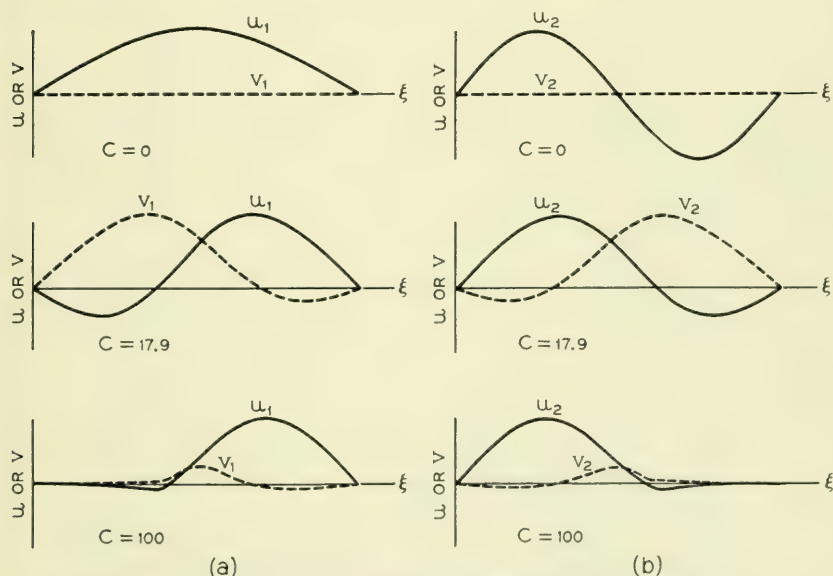


Fig. 23—Real and imaginary parts of the first two eigenfunctions, $w_1 = u_1 + iv_1$ and $w_2 = u_2 + iv_2$, for the nonuniform stack of Fig. 22.

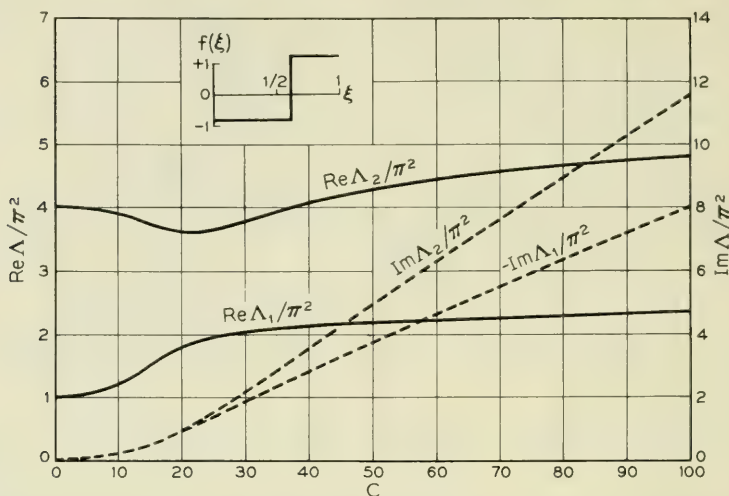


Fig. 24—Real and imaginary parts of Λ_1/π^2 and Λ_2/π^2 for a nonuniform stack whose average properties are constant except at a single, unsymmetric discontinuity.

$$\Lambda \approx \frac{\pi^2}{(1 - \xi_0)^2} \left[1 - 4 \sqrt{\frac{\xi_0}{(1 - \xi_0)C}} \right] + i \left[\frac{C}{2(1 - \xi_0)} - \frac{4\pi^2}{(1 - \xi_0)^2} \sqrt{\frac{\xi_0}{(1 - \xi_0)C}} \right]. \quad (550)$$

Numerical values of Λ_1 were found for the case $\xi_0 = \frac{1}{2}$ on the analog computer and refined afterward by desk computation. The real and imaginary parts of Λ_1/π^2 are plotted over the range $0 \leq C \leq 100$ in Fig. 25, and the corresponding eigenfunction $w_1(\xi) = u_1(\xi) + iw_1(\xi)$ is shown in Fig. 26 for $C = 0, 20$, and 100 .

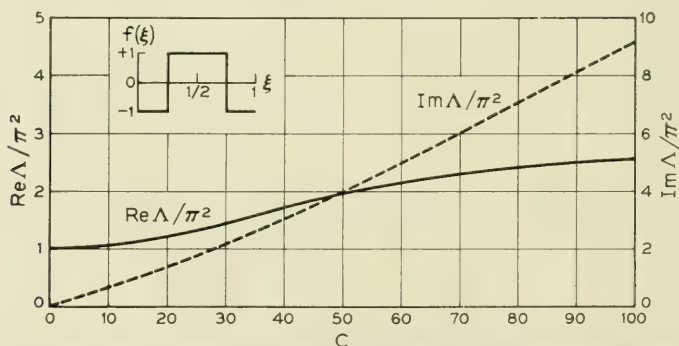


Fig. 25—Real and imaginary parts of Λ_1/π^2 for a nonuniform stack whose average properties vary as a symmetric rectangular step.

(iii) $f(\xi)$ a linear function. Let

$$f(\xi) = 2\xi - 1, \quad 0 \leq \xi \leq 1, \quad (551)$$

so that $f(\xi)$ is a linear function varying from -1 at $\xi = 0$ to $+1$ at $\xi = 1$. Then equation (532) becomes

$$d^2w/d\xi^2 + [(\Lambda + iC) - 2iC\xi]w(\xi) = 0, \quad (552)$$

which, by the change of variable

$$\tau = \frac{2iC\xi - (\Lambda + iC)}{(2C)^{2/3}}, \quad (553)$$

is transformed into Stokes' equation,

$$d^2w/d\tau^2 + \tau w = 0. \quad (554)$$

The general solution of this equation may be written in the form

$$w = Ah_1(\tau) + Bh_2(\tau), \quad (555)$$

where h_1 and h_2 are the pair of independent solutions of Stokes' equation which have been tabulated for complex arguments by the Computation Laboratory of Harvard University.²⁷ (The solution may also be ex-

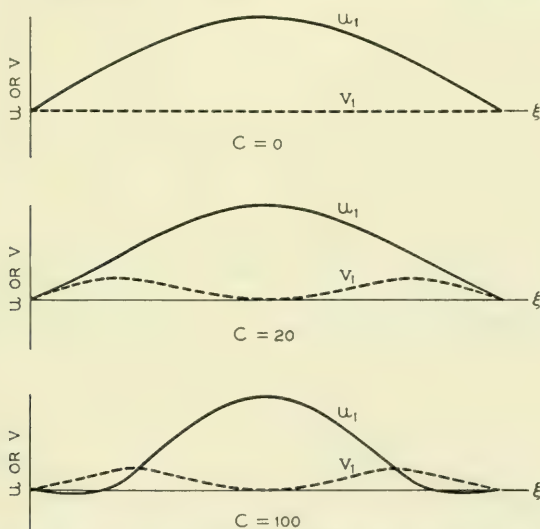


Fig. 26—Real and imaginary parts of the first eigenfunction, $w_1 = u_1 + iv_1$, for the nonuniform stack of Fig. 25.

²⁷ *Tables of the Modified Hankel Functions of Order One-Third and of Their Derivatives*, Harvard University Press, Cambridge, Mass., 1945.

pressed, though less conveniently, in terms of Bessel functions of order one-third.) It is easy to show that the boundary conditions (533) at $\xi = 0$ and $\xi = 1$ require

$$\begin{aligned} Ah_1(\tau_1) + Bh_2(\tau_1) &= 0, \\ Ah_1(\tau_2) + Bh_2(\tau_2) &= 0, \end{aligned} \quad (556)$$

where

$$\tau_1 = -\frac{(\Lambda + iC)}{(2C)^{\frac{1}{3}}}, \quad \tau_2 = -\frac{(\Lambda - iC)}{(2C)^{\frac{1}{3}}}. \quad (557)$$

Equations (556) will be consistent if

$$h_1(\tau_1)h_2(\tau_2) - h_1(\tau_2)h_2(\tau_1) = 0; \quad (558)$$

and this is the relation which must be satisfied by the eigenvalues $\Lambda_1, \Lambda_2, \Lambda_3, \dots$, for any given value of C .

Approximate values of Λ_1 and Λ_2 have been found using the analog computer for the range $0 \leq C \leq 100$, with spot checks by numerical solution of equation (558); and Λ_1/π^2 and Λ_2/π^2 are plotted in Fig. 27. The eigenfunctions are qualitatively similar to those shown in Fig. 23 for the stack with a symmetric discontinuity. As in the symmetric example in case (i) above, we find that for small positive C , Λ_1 is real and greater than π^2 , while Λ_2 is real and less than $4\pi^2$. The two eigenvalues coincide at

$$C \approx 49. \quad \Lambda_1 = \Lambda_2 \approx 29. \quad (559)$$

For larger values of C , Λ_1 and Λ_2 are complex conjugates. Their asymp-

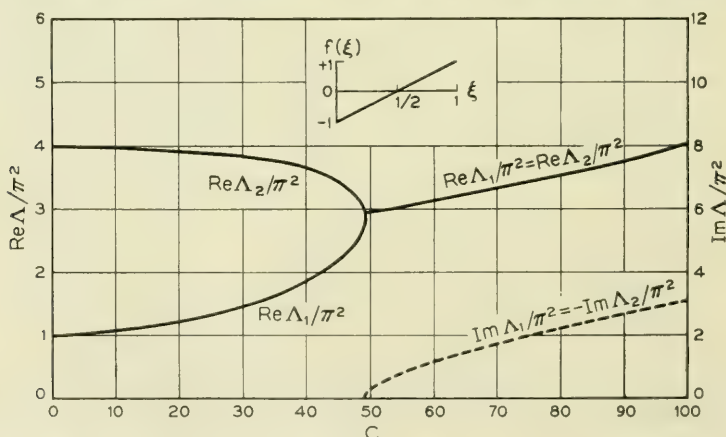


Fig. 27—Real and imaginary parts of Λ_1/π^2 and Λ_2/π^2 for a nonuniform stack whose average properties vary linearly across the stack.

otic forms as $C \rightarrow \infty$ may be deduced by considering the behavior of $h_1(\tau)$ and $h_2(\tau)$ for large arguments, and are

$$\Lambda_1 = \Lambda_2^* \approx 1.169(2C)^{\frac{2}{3}} + i[C - 2.025(2C)^{\frac{2}{3}}]. \quad (560)$$

The magnitudes of both the real and imaginary parts of Λ_1 and Λ_2 thus increase indefinitely with C .

(iv) $f(\xi)$ a sinusoidal function. Let

$$f(\xi) = -\cos 2\nu\pi\xi, \quad 0 \leq \xi \leq 1, \quad (561)$$

where $\nu = \frac{1}{2}, 1, 2, 3, 4, \dots$, so that $f(\xi)$ goes through ν complete cycles in $0 \leq \xi \leq 1$. Then equation (532) reads

$$d^2w/d\xi^2 + [\Lambda + iC \cos 2\nu\pi\xi]w(\xi) = 0. \quad (562)$$

If we make the transformations

$$\begin{aligned} \tau &= \nu\pi\xi, \\ W(\tau) &= w(\xi), \\ \lambda &= \Lambda/\nu^2\pi^2, \\ \vartheta &= -iC/2\nu^2\pi^2, \end{aligned} \quad (563)$$

we get

$$d^2W/d\tau^2 + [\lambda - 2\vartheta \cos 2\tau]W(\tau) = 0, \quad (564)$$

and the boundary conditions (533) become

$$W(0) = W(\nu\pi) = 0. \quad (565)$$

Equation (564) is one of the standard forms of Mathieu's equation. We are interested in solutions which are periodic with period 2 in ξ , and which approach the form $\sin m\pi\xi$ when $C \rightarrow 0$. In terms of τ and ϑ , the function corresponding to the m th mode in the Clogston line must reduce to the form

$$W(\tau) \xrightarrow[\vartheta \rightarrow 0]{} \sin \frac{m}{\nu} \tau. \quad (566)$$

For any value of ϑ , this function may be denoted by²⁸

$$W(\tau) = \text{se}_{m/\nu}(\tau, \vartheta). \quad (567)$$

²⁸ See N. W. McLachlan, *Theory and Application of Mathieu Functions*, Oxford, 1947, pp. 10-25, especially p. 13 and p. 19. In this reference a or b corresponds to our λ , q to our ϑ , and ν to our m/ν .

In our problem ϑ is (negative) imaginary and m/ν may be an integer or a rational fraction. For any given ϑ and m/ν the conditions (565) together with the limiting form (566) determine an eigenvalue λ , and hence by (563) determine Λ ; but only a small amount of work has been published on the eigenvalues of Mathieu functions with imaginary parameter or of fractional order. We shall look at some special cases.

$\nu = \frac{1}{2}$. The function $f(\xi)$ is one-half cycle of a cosine curve which varies from -1 to $+1$; we expect results similar to those found for the symmetric discontinuity of case (i) and the linear variation of case (iii). The eigenfunctions of the first two modes ($m = 1$ and $m = 2$) are $se_2(\tau, \vartheta)$ and $se_4(\tau, \vartheta)$. The eigenvalues of these two functions for purely imaginary ϑ have been computed by Mulholland and Goldstein²⁹ out to a point which corresponds to $C = 8\pi^2$, and an asymptotic formula is given for larger values of C . The values of Λ_1/π^2 and Λ_2/π^2 are plotted for $0 \leq C \leq 100$ in Fig. 28; the corresponding eigenfunctions resemble those shown in Fig. 23 for the stack with a symmetric discontinuity. Again we find that Λ_1 and Λ_2 are real for small positive C , equal for a particular value of C , and conjugate complex for larger C . The leading terms of the asymptotic formula are, in our notation,

$$\begin{aligned} \Lambda_1 = \Lambda_2^* \approx [4.7124C^{\frac{1}{2}} - 3.0842 - 1.0901C^{-\frac{1}{2}} - \dots] \\ + i[C - 4.7124C^{\frac{1}{2}} - 1.0901C^{-\frac{1}{2}} - \dots]. \end{aligned} \quad (568)$$

$\nu = 1$. Here $f(\xi)$ is one full cycle of a cosine function, varying from -1

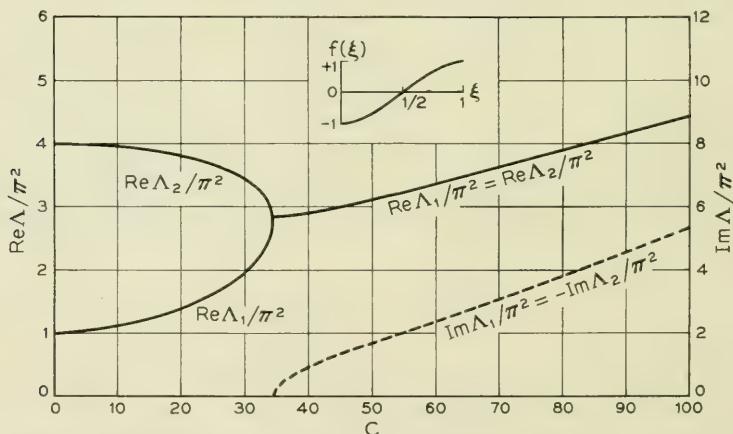


Fig. 28—Real and imaginary parts of Λ_1/π^2 and Λ_2/π^2 for a nonuniform stack whose average properties vary as one-half cycle of a cosine function across the stack.

²⁹ H. P. Mulholland and S. Goldstein, *Phil. Mag.* (7), **8**, 834 (1929). In this reference 4α or 4β corresponds to our λ and $8q$ to our ϑ .

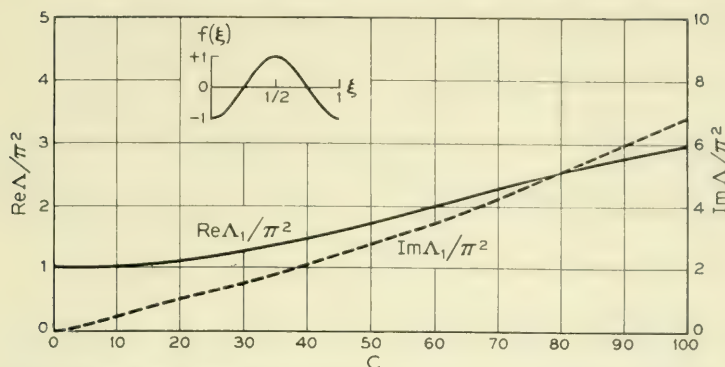


Fig. 29—Real and imaginary parts of Λ_1/π^2 for a nonuniform stack whose average properties vary as one cycle of a cosine function across the stack.

to $+1$ and back to -1 . The eigenfunction of the lowest mode ($m = 1$) is $se_1(\tau, \vartheta)$, and the values of Λ_1 may be obtained from Reference 29 for ten equally spaced values of C out to $C = 32\pi^2$. Since our ϑ is negative imaginary, in the notation of this reference we have $\Lambda_1 = 4\pi^2\beta_1^*$. Approximate values of Λ_1/π^2 obtained on the analog computer for C at smaller intervals in the range $0 \leq C \leq 100$ are plotted in Fig. 29; and the eigenfunctions are similar to those shown in Fig. 26 for the symmetric rectangular step. The leading terms of the asymptotic formula for Λ_1 when C is large are as follows:

$$\begin{aligned} \Lambda_1 \approx [3.1416C^{\frac{1}{2}} - 2.4674 - 0.9689C^{-\frac{1}{2}} - \dots] \\ + i[C - 3.1416C^{\frac{1}{2}} - 0.9689C^{-\frac{1}{2}} - \dots]. \end{aligned} \quad (569)$$

$\nu = 3$. Now $f(\xi)$ is a three-cycle cosine function and the lowest mode corresponds to $se_3(\tau, \vartheta)$. Approximate values of Λ_1/π^2 for $0 \leq C \leq 100$ were obtained on the analog computer and are plotted in Fig. 30; the eigenfunctions are shown in Fig. 31 for $C = 0$ and $C = 100$.

$\nu \gg 1$. For a ν -cycle cosine variation, the lowest eigenfunction is $se_{1/\nu}(\tau, \vartheta)$, and for the lowest eigenvalue there is an approximate formula given by McLachlan.³⁰ Incidentally this formula predicts no imaginary part for Λ_1 if ϑ is purely imaginary and $\nu > 1$, which agrees approximately with the results of our analog computations for $\nu = 3$; we found the imaginary part of Λ_1 to be only about 1 per cent of the real part even for $C = 100$. If C is fixed, one expects that as $\nu \rightarrow \infty$ the effects of the rapid fluctuations in $f(\xi)$ will average out, so that Λ_1 will ultimately approach

³⁰ Reference 28, p. 20, equation (6), where a corresponds to our λ_1 , q to our ϑ , and ν to our $1/\nu$. McLachlan's formula was ostensibly derived for real q , but the derivation appears equally valid for complex q .

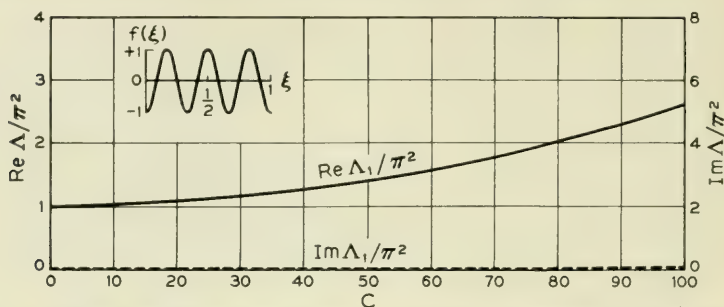


Fig. 30—Real and imaginary parts of Λ_1/π^2 for a nonuniform stack whose average properties vary as three cycles of a cosine function across the stack.

the value π^2 appropriate to a uniform stack. McLachlan's formula shows that this is indeed the case; in our notation, the leading terms give

$$\Lambda \approx \pi^2 + \frac{C^2}{8\nu^2\pi^2} = \pi^2 \left[1 + \frac{C^2}{8\nu^2\pi^4} \right], \quad (570)$$

assuming of course that the second term is reasonably small compared to the first.

This concludes our discussion of special types of nonuniformity. We shall now attempt to get an idea of what the numerical results mean in terms of the practical requirements on stack uniformity in a laminated transmission line which is expected to show a specified reduction in attenuation constant below a conventional line of the same dimensions. For this purpose we shall compare a plane Clogston 2 line having infinitesimally thin layers with a plane air-filled line of the same width a , bounded by electrically thick solid conductors.

At frequencies for which the conductor thickness of the "standard"

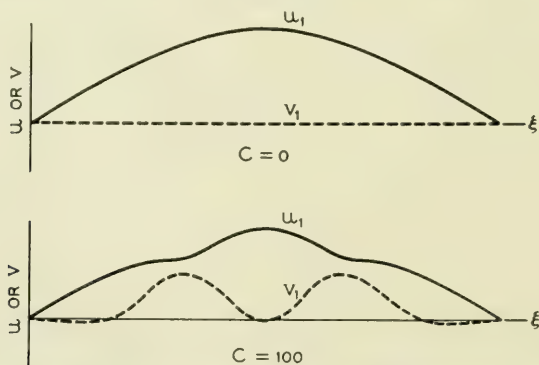


Fig. 31—Real and imaginary parts of the first eigenfunction, $w_1 = u_1 + iv_1$, for the nonuniform stack of Fig. 30.

air-filled line is great compared to the skin depth δ_1 , its attenuation constant α_s is given by equation (25), namely

$$\alpha_s = 1/\eta_v g_1 \delta_1 a, \quad (571)$$

where η_v is the intrinsic impedance of free space. By equation (535), the attenuation constant α_c of the lowest mode in a plane Clogston 2 with infinitesimally thin layers is

$$\alpha_c = \text{Re} \frac{\Lambda_1}{2\sqrt{\bar{\mu}_0/\bar{\epsilon}_0} \bar{g}_0 a^2}. \quad (572)$$

If we assume nonmagnetic materials and put in the optimum value of θ , namely $\theta = 2/3$, we obtain for a uniform stack with $\Lambda_1 = \pi^2$,

$$\alpha_{c0} = \frac{12.82\sqrt{\epsilon_{2r}}}{\eta_v g_1 a^2}, \quad (573)$$

where ϵ_{2r} is the relative dielectric constant of the insulating layers.

The attenuation constant of the conventional line is proportional to the square root of frequency, whereas the attenuation constant of the uniform Clogston 2 is independent of frequency up to some frequency at which the effect of finite lamina thickness begins to be appreciable. If we confine ourselves to the low-frequency, flat attenuation region, and denote the ratio of attenuation constants by r , then from (571) and (573),

$$r = \alpha_{c0}/\alpha_s = 12.82\sqrt{\epsilon_{2r}} \delta_1/a, \quad (574)$$

and the crossover frequency above which the uniform Clogston line is better than the conventional line occurs when

$$a/\delta_1 = 12.82\sqrt{\epsilon_{2r}}. \quad (575)$$

In the following numerical example we shall assume polyethylene insulating layers, with

$$\epsilon_{2r} = 2.26, \quad (576)$$

so that (574) becomes

$$r = 19.27\delta_1/a. \quad (577)$$

If the stack in a Clogston line is not uniform, then regardless of the thinness of the layers the attenuation constant will no longer be independent of frequency, but will increase with frequency at a rate depending on the nature and the magnitude of the nonuniformity. Since from equation (572) the attenuation constant is proportional to $\text{Re } \Lambda_1$,

while from equation (529) or (530), C is proportional to frequency for a given stack, we see that our plots of $\text{Re } \Lambda_1/\pi^2$ versus C need only the introduction of appropriate scale factors to read directly the variation of attenuation with frequency due to nonuniformity in the stack. Although a nonuniform Clogston line may still be better under some conditions than a conventional line of the same size, the crossover frequency will be higher and the improvement at any given frequency will be less than if the stack were uniform.

Among the various interpretations which may be given to our numerical results, we shall consider here only the following: Suppose we have a plane Clogston 2 line which, if it were perfectly uniform, would have an attenuation constant smaller, at a certain frequency, than the attenuation constant of the corresponding conventional line by a given factor, say one-half, one-fifth, or one-tenth. For these particular attenuation reduction factors the ratio of a to δ_1 may be calculated from (574), or from (577) if the insulation is polyethylene. The question is: What variation in $\bar{\epsilon}$ across the stack is permissible if we are willing to have the actual attenuation constant of the Clogston line be double its ideal value; in other words, if we will settle for attenuation reduction factors of unity (no improvement), two-fifths, or one-fifth instead of the ideal values one-half, one-fifth, or one-tenth?

To answer this question for any particular type of nonuniformity, we have only to find, from the plot of $\text{Re } \Lambda_1/\pi^2$ versus C , the value of C for which $\text{Re } \Lambda_1/\pi^2 = 2$. Then the fractional difference between the maximum and minimum values of $\bar{\epsilon}$ corresponding to this value of C is given by equations (530) and (531) to be

$$\frac{\bar{\epsilon}_{\max} - \bar{\epsilon}_{\min}}{\bar{\epsilon}_0} = \frac{3C\delta_1^2}{4a^2} (f_{\max} - f_{\min}), \quad (578)$$

where we have taken $\theta_0 = 2/3$, and f_{\max} and f_{\min} are the extreme values of the function $f(\xi)$ which describes the type of nonuniformity being considered.

The special types of nonuniformity which have been studied above fall roughly into three different classes. In four of the cases, namely the symmetric and unsymmetric single discontinuities, the linear variation, and the half-cycle cosine variation, the function $f(\xi)$ varies monotonically from one side of the stack to the other. In the symmetric rectangular step and the one-cycle cosine variation, $f(\xi)$ oscillates from one extreme value to the other and back again, while in the three-cycle cosine variation, $f(\xi)$ exhibits three complete oscillations across the stack. The following table

shows the permissible total variation in $\bar{\epsilon}$ for each of these types of non-uniformity.

Case	C	$\frac{\bar{\epsilon}_{\max} - \bar{\epsilon}_{\min}}{\bar{\epsilon}_0}$		
		$r = \frac{1}{2}$	$r = \frac{1}{5}$	$r = \frac{1}{10}$
Symmetric discontinuity.....	16.5	0.0167	0.0027	0.0007
Unsymmetric discontinuity.....	28.0	0.0295	0.0047	0.0012
Linear.....	42.6	0.0430	0.0069	0.0017
Half-cycle cosine.....	29.5	0.0298	0.0048	0.0012
Rectangular step.....	53.0	0.0535	0.0086	0.0021
One-cycle cosine.....	59.8	0.0604	0.0097	0.0024
Three-cycle cosine.....	78.9	0.0797	0.0128	0.0032

It would be easy to construct a similar table for any other values of the attenuation ratio r , and for any specified degradation due to nonuniformity. It is, however, already obvious that the greater the improvement for which one strives, that is, the smaller the ratio r , the more stringent will be the requirement on $(\bar{\epsilon}_{\max} - \bar{\epsilon}_{\min})/\bar{\epsilon}_0$; in fact, the permissible value of this quantity is proportional to r^2 . In any practical case the value of $\bar{\epsilon}$ will have to be controlled against long-range variations within a fraction of a per cent, and if attenuation reduction factors of the order of one-fifth or one-tenth are contemplated, the variations probably cannot exceed a few hundredths of a per cent. It also appears that a steady increase or decrease in the value of $\bar{\epsilon}$ across the stack will be the most serious type of nonuniformity, since the effects of very rapid fluctuations will tend to average out.

Clearly the nonuniform laminated transmission lines which we have been considering in this section are very highly idealized, even if we disregard the geometrical differences between plane and coaxial structures. Any real Clogston cable will be built up of layers of finite thickness with unavoidable random fluctuations from layer to layer, superimposed on slower variations in the average properties of the layers from one side of the stack to the other. The thickness of an individual layer will also vary more or less in both directions parallel to the layer, so that the properties of the stack will be functions of the coordinates ϕ and z as well as of ρ . A few qualitative remarks are in order concerning these neglected effects.

The effect of finite lamina thickness in a nonuniform stack can be calculated, by the method employed in Section XI for a uniform coaxial stack, if we make the plausible assumption that the macroscopic current distribution remains the same as for infinitesimally thin layers. The results

will certainly be qualitatively the same for uniform and slightly nonuniform stacks, so long as the nonuniformity does not seriously distort the field pattern of the operating mode.

Some idea of how the effects of rapid random fluctuations in the average properties of the stack may be expected to average out is given by equation (570), which assumes for the function $f(\xi)$ a ν -cycle cosine variation across the stack. As a numerical example, suppose that with this variation of $\bar{\epsilon}$ we have

$$\frac{\bar{\epsilon}_{\max} - \bar{\epsilon}_{\min}}{\bar{\epsilon}_0} = 0.01 \quad (579)$$

in a line designed to give an attenuation reduction ratio of

$$r = 1/10. \quad (580)$$

Assuming polyethylene insulating layers, we have for this line

$$\delta_1/a = 1/192.7 = 0.00519, \quad (581)$$

and from (578) the corresponding value of C is

$$C = 247.6. \quad (582)$$

The value of ν for which the relative increase in $\text{Re } \Lambda_1$ due to the fluctuations is, say, one-quarter is given by

$$\frac{C^2}{8\nu^2\pi^4} = \frac{1}{4}, \quad \nu = \frac{C}{\sqrt{2}\pi^2} = 17.7. \quad (583)$$

Thus a 1 per cent fluctuation in $\bar{\epsilon}$, repeated at intervals of about one-eighteenth of the stack width, will cause only a 25 per cent increase in attenuation, even for a Clogston line which is designed to have only one-tenth of the attenuation constant of a conventional line of the same size.

Finally there is the question of the effects of variations in the average properties of the stack in both directions parallel to the layers. Mathematical analysis of even a simple case of longitudinal variation would be much more difficult than what has been done here; yet on physical grounds it seems very likely that such variations will add an appreciable amount to the total attenuation of the line. If we consider two cross sections of a laminated cable separated by a certain distance and having different transverse nonuniformities, the field pattern of the lowest mode will be different at the two cross sections, and so in traversing the intervening distance the power will be partly reflected and partly converted to higher modes with higher attenuation constants. The reflected or mode converted power will be at least partly lost, with a consequent increase in the overall at-

tenuation of the cable. Hence the estimate of the increase in attenuation which one gets from the present analysis, considering only the variations transverse to the layers at an average cross section, is certain to be optimistic in that it neglects completely the effects of variations in other directions.

XIII. DIELECTRIC AND MAGNETIC LOSSES IN CLOGSTON 2 LINES

To discuss dielectric and magnetic losses in Clogston 2 lines we may take the electrical constants of the conducting and insulating layers to be complex; thus

$$\begin{aligned}\mu_1 &= \mu'_1 - i\mu''_1 = \mu'_1(1 - i \tan \zeta_1), \\ \mu_2 &= \mu'_2 - i\mu''_2 = \mu'_2(1 - i \tan \zeta_2), \\ \epsilon_2 &= \epsilon'_2 - i\epsilon''_2 = \epsilon'_2(1 - i \tan \phi_2).\end{aligned}\tag{584}$$

Almost all of the equations of the preceding sections, except of course those which involve explicit separation of real and imaginary parts, remain valid when we introduce complex values of μ_1 , μ_2 , and ϵ_2 . In particular the propagation constant of the p th mode in a Clogston 2 with infinitesimally thin laminae and high-impedance walls is given, as in Sections VIII through X, by

$$\gamma^2 = -\omega^2 \bar{\mu} \bar{\epsilon} + (i\omega \bar{\epsilon} / \bar{g}) \chi_p^2, \tag{585}$$

where

$$\chi_p = p\pi/a \tag{586}$$

for a parallel-plane line, and χ_p is the p th root of

$$J_1(\chi a)N_1(\chi b) - J_1(\chi b)N_1(\chi a) = 0 \tag{587}$$

for a coaxial line. Taking the square root of the right side of (585) by the binomial theorem, we have

$$\gamma = i\omega \sqrt{\bar{\mu} \bar{\epsilon}} + \frac{\chi_p^2}{2\sqrt{\bar{\mu} \bar{\epsilon}} \bar{g}}. \tag{588}$$

In the presence of dielectric and/or magnetic dissipation, we write, as in Section VII,

$$\begin{aligned}\bar{\epsilon} &= \bar{\epsilon}' - i\bar{\epsilon}'' = [\epsilon'_2/(1 - \theta)] - i[\epsilon''_2/(1 - \theta)], \\ \bar{\mu} &= \bar{\mu}' - i\bar{\mu}'' = [\theta\mu'_1 + (1 - \theta)\mu'_2] - i[\theta\mu''_1 + (1 - \theta)\mu''_2].\end{aligned}\tag{589}$$

Then the term $i\omega\sqrt{\bar{\mu}\bar{\epsilon}}$ in (588) has a small real part, namely

$$\alpha_d = \frac{1}{2}\omega\sqrt{\bar{\mu}'\bar{\epsilon}'}(\tan\phi_2 + \tan\bar{\xi}), \quad (590)$$

where

$$\tan\bar{\xi} = \frac{\bar{\mu}''}{\bar{\mu}'} = \frac{\theta\mu_1'' + (1-\theta)\mu_2''}{\theta\mu_1' + (1-\theta)\mu_2'}; \quad (591)$$

and α_d is the part of the attenuation constant which is due to dielectric and magnetic losses. If there were no dielectric or magnetic dissipation, the second term on the right side of (588) would be purely real and would represent the attenuation due to ohmic losses in the conducting layers. We neglect the small change in this term when $\bar{\mu}$ and $\bar{\epsilon}$ are complex, and thus as usual regard the metal losses, the dielectric losses, and the magnetic losses as additive.

We observe that α_d is the same for both plane and coaxial lines, and is also independent of the mode number p . Although derived here for the case of infinitesimally thin laminae, the same expression may be used for lines with finite laminae, so long as the conducting layers are moderately thin compared to the skin depth. The dielectric and magnetic losses do not depend on the overall dimensions of the transmission line, but are directly proportional to frequency provided that the loss tangents do not vary with frequency.

If it should be necessary to calculate the dielectric and magnetic losses in a partially filled Clogston line where the dissipation factor of the main dielectric is markedly different from the dissipation factor of the stacks, α_d may be obtained, using the method described in Section VII, as half the ratio of dissipated power per unit length to transmitted power. In this calculation we may use the field components given in Sections IX and X for the various modes in partially filled lines.

ACKNOWLEDGMENTS

Many people with whom I have discussed the theoretical and practical aspects of the laminated transmission line problem at various times have offered comments and suggestions which are reflected in this paper. I have especially to express my appreciation to A. M. Clogston, H. S. Black, and J. G. Kreer, Jr., for stimulating and helpful discussions.

My thanks are also extended to Mrs. M. F. Shearer, Mrs. D. R. Fursdon, and Miss R. A. Weiss for the extensive numerical computations which they carried out in connection with this study, and to Miss D. T. Angell for preparation of the curves and diagrams.

APPENDIX II

OPTIMUM PROPORTIONS FOR HEAVILY LOADED CLOGSTON CABLES

We wish to find the lowest root χ_1 of the equation

$$\frac{1}{\chi \rho_1} \frac{J_1(\chi a) N_0(\chi \rho_1) - N_1(\chi a) J_0(\chi \rho_1)}{J_1(\chi a) N_1(\chi \rho_1) - N_1(\chi a) J_1(\chi \rho_1)} + \frac{1}{\chi \rho_2} \frac{J_1(\chi b) N_0(\chi \rho_2) - N_1(\chi b) J_0(\chi \rho_2)}{J_1(\chi \rho_2) N_1(\chi b) - N_1(\chi \rho_2) J_1(\chi b)} = \frac{\mu_0}{\bar{\mu}} \log \frac{\rho_2}{\rho_1}, \quad (\text{A9})$$

where b is fixed and $\mu_0/\bar{\mu} \gg 1$, and to minimize this root as a function of a , ρ_1 , and ρ_2 .

Since we expect χ_1 to approach zero as $\mu_0/\bar{\mu}$ approaches infinity, we shall replace the Bessel functions appearing in (A9) by their approximate values for small argument, namely

$$\begin{aligned} J_0(x) &\approx 1, \\ J_1(x) &\approx \frac{1}{2}x, \\ N_0(x) &\approx \frac{2}{\pi} \log 0.8905x, \\ N_1(x) &\approx -\frac{2}{\pi x}, \end{aligned} \quad (\text{A10})$$

for $|x| \ll 1$. Then the equation becomes, approximately,

$$\frac{1}{\chi_1 \rho_1} \frac{1/\chi_1 a}{\frac{1}{2}(-a/\rho_1 + \rho_1/a)} + \frac{1}{\chi_1 \rho_2} \frac{1/\chi_1 b}{\frac{1}{2}(-\rho_2/b + b/\rho_2)} = \frac{\mu_0}{\bar{\mu}} \log \frac{\rho_2}{\rho_1}, \quad (\text{A11})$$

which may be solved for χ_1^2 to yield

$$\chi_1^2 = \frac{2\bar{\mu}}{\mu_0 \log(\rho_2/\rho_1)} \left[\frac{1}{\rho_1^2 - a^2} + \frac{1}{b^2 - \rho_2^2} \right]. \quad (\text{A12})$$

By inspection χ_1^2 will be a minimum, considered as a function of a , when $a = 0$. Setting $a = 0$ and then equating to zero the partial derivatives of χ_1^2 with respect to ρ_1 and ρ_2 , we get the pair of equations

$$\begin{aligned} \frac{1}{\rho_1 [\log(\rho_2/\rho_1)]^2} \left[\frac{1}{\rho_1^2} + \frac{1}{b^2 - \rho_2^2} \right] - \frac{2}{\rho_1^3 \log(\rho_2/\rho_1)} &= 0, \\ -\frac{1}{\rho_2 [\log(\rho_2/\rho_1)]^2} \left[\frac{1}{\rho_1^2} + \frac{1}{b^2 - \rho_2^2} \right] + \frac{2\rho_2}{(b^2 - \rho_2^2)^2 \log(\rho_2/\rho_1)} &= 0, \end{aligned} \quad (\text{A13})$$

which yield, on rearrangement,

$$\begin{aligned} \frac{1}{\log (\rho_2/\rho_1)} \left[\frac{1}{\rho_1^2} + \frac{1}{b^2 - \rho_2^2} \right] &= \frac{2}{\rho_1^2}, \\ \frac{1}{\log (\rho_2/\rho_1)} \left[\frac{1}{\rho_1^2} + \frac{1}{b^2 - \rho_2^2} \right] &= \frac{2\rho_2^2}{(b^2 - \rho_2^2)^2}. \end{aligned} \quad (\text{A14})$$

Subtracting the second of equations (A14) from the first and solving the resulting equation for ρ_1 , we get

$$\rho_1 = \frac{b^2 - \rho_2^2}{\rho_2} = \frac{b^2}{\rho_2} - \rho_2, \quad (\text{A15})$$

whence, eliminating ρ_1 from the first of (A14),

$$\log \frac{\rho_2^2}{b^2 - \rho_2^2} = \frac{b^2}{2\rho_2^2}. \quad (\text{A16})$$

Numerical solution of (A16) gives

$$\rho_2^2/b^2 = 0.67674; \quad (\text{A17})$$

and on making use of (A15) we obtain finally

$$\begin{aligned} \rho_1 &= 0.39296b, \\ \rho_2 &= 0.82264b. \end{aligned} \quad (\text{A18})$$

Substituting these values, with $a = 0$, into (A12), we get for the minimum value of χ_1^2 , when $\mu_0/\bar{\mu} \gg 1$,

$$\chi_1^2 = \frac{25.905\bar{\mu}}{\mu b^2}. \quad (\text{A19})$$

APPENDIX III

POWER DISSIPATION IN A HOLLOW CONDUCTING CYLINDER

Consider a hollow cylinder of inner radius ρ_1 , outer radius ρ_2 , and high conductivity g_1 . Denote the total current flowing in the positive z -direction inside the radius ρ_1 by I_1 and the total current inside the radius ρ_2 by I_2 ; then the current carried by the conducting cylinder is just $I_2 - I_1$, and the net return current outside the cylinder is $-I_2$. We assume the current distribution to be independent of the coordinate angle ϕ , but the radial distribution of the currents inside and outside the given cylinder is of no importance.

General expressions for the field components in the conducting cylinder

are given by equations (33) of Section II, which read

$$\begin{aligned} H_\phi &= AI_1(\sigma_1\rho) + BK_1(\sigma_1\rho), \\ E_\rho &= \frac{\gamma}{g_1} [AI_1(\sigma_1\rho) + BK_1(\sigma_1\rho)], \\ E_z &= \eta_1[AI_0(\sigma_1\rho) - BK_0(\sigma_1\rho)], \end{aligned} \quad (\text{A20})$$

provided that we drop the propagation factor $e^{-\gamma z}$ and make the usual approximations

$$g_1/\omega\epsilon_1 \gg 1, \quad \kappa_1 \approx \sigma_1, \quad \kappa_1/g_1 \approx \eta_1, \quad (\text{A21})$$

for a good conductor. The constants A and B are determined by the boundary conditions

$$H_\phi(\rho_1) = I_1/2\pi\rho_1, \quad H_\phi(\rho_2) = I_2/2\pi\rho_2, \quad (\text{A22})$$

which follow directly from Ampere's circuital law. We find without difficulty

$$\begin{aligned} A &= \frac{(I_2/2\pi\rho_2)K_1(\sigma_1\rho_1) - (I_1/2\pi\rho_1)K_1(\sigma_1\rho_2)}{K_1(\sigma_1\rho_1)I_1(\sigma_1\rho_2) - K_1(\sigma_1\rho_2)I_1(\sigma_1\rho_1)}, \\ B &= \frac{(I_1/2\pi\rho_1)I_1(\sigma_1\rho_2) - (I_2/2\pi\rho_2)I_1(\sigma_1\rho_1)}{K_1(\sigma_1\rho_1)I_1(\sigma_1\rho_2) - K_1(\sigma_1\rho_2)I_1(\sigma_1\rho_1)}. \end{aligned} \quad (\text{A23})$$

The average power dissipated in the conducting cylinder is equal to one-half the real part of the inward normal flux of the complex Poynting vector $\mathbf{E} \times \mathbf{H}^*$. For the average power P dissipated per unit length we have

$$\begin{aligned} P &= \text{Re } \frac{1}{2}[2\pi\rho_2 E_z(\rho_2)H_\phi^*(\rho_2) - 2\pi\rho_1 E_z(\rho_1)H_\phi^*(\rho_1)] \\ &= \text{Re } \frac{1}{2}[E_z(\rho_2)I_2^* - E_z(\rho_1)I_1^*] \\ &= \text{Re } \frac{\frac{1}{2}\eta_1}{(K_{11}I_{12} - K_{12}I_{11})} \left[\frac{I_2 I_2^*}{2\pi\rho_2} (K_{11}I_{02} + K_{02}I_{11}) \right. \\ &\quad \left. - \frac{(I_1 I_2^* + I_1^* I_2)}{2\pi\sigma_1\rho_1\rho_2} + \frac{I_1 I_1^*}{2\pi\rho_2} (K_{01}I_{12} + K_{12}I_{01}) \right], \end{aligned} \quad (\text{A24})$$

where

$$I_{rs} = I_r(\sigma_1\rho_s), \quad K_{rs} = K_r(\sigma_1\rho_s). \quad (\text{A25})$$

The combinations of Bessel functions appearing in (A24) are just those for which we gave approximate expressions in equations (A8) of Appendix I, assuming the thickness $t_1 (= \rho_2 - \rho_1)$ of the conducting cylinder to be small compared to ρ_1 . Substituting these approximations into (A24) and

rearranging, we get

$$P = \text{Re} \frac{1}{4\pi g_1 t_1} \left\{ \frac{I_2 I_2^*}{\rho_2} \left[\sigma_1 t_1 \coth \sigma_1 t_1 + \frac{t_1}{2\rho_1} \right] - \frac{(I_1 I_2^* + I_1^* I_2)}{\rho_1} \left[1 - \frac{t_1}{2\rho_1} \right] \sigma_1 t_1 \text{csch } \sigma_1 t_1 + \frac{I_1 I_1^*}{\rho_1} \left[\sigma_1 t_1 \coth \sigma_1 t_1 - \frac{t_1}{2\rho_1} \right] \right\}, \quad (\text{A26})$$

up to first order in t_1/ρ_1 . If we set

$$\sigma_1 = (1 + i)/\delta_1, \quad (\text{A27})$$

then on expanding the right side of (A26) in powers of t_1/δ_1 up to the fourth, we obtain

$$P = \frac{1}{4\pi g_1 t_1} \left\{ \frac{|I_2 - I_1|^2}{\sqrt{\rho_1 \rho_2}} + \frac{4t_1^4}{\delta_1^4} \left[\frac{I_2 I_2^*}{45\rho_2} + \frac{7(I_1 I_2^* + I_1^* I_2)}{360\sqrt{\rho_1 \rho_2}} + \frac{I_1 I_1^*}{45\rho_1} \right] \right\}, \quad (\text{A28})$$

where we have approximated $\rho_1(1 + t_1/2\rho_1)$ by $\sqrt{\rho_1 \rho_2}$ in the interest of symmetry.

Now writing ΔP_j for P , ΔI_j for $I_2 - I_1$, ρ_{j-1} for ρ_1 , and I_{j-1} for I_1 , and neglecting curvature corrections of the order of t_1/ρ_1 entirely, we have, on setting I_1 and I_2 both equal to I_{j-1} in the coefficient of $(t_1/\delta_1)^4$, the approximate relation

$$\Delta P_j = \frac{1}{4\pi g_1 t_1 \rho_{j-1}} \left[|\Delta I_j|^2 + \frac{t_1^4}{3\delta_1^4} |I_{j-1}|^2 \right], \quad (\text{A29})$$

which is just equation (472) of Section XI.

Transistors in Switching Circuits

By A. EUGENE ANDERSON

(Manuscript received August 1, 1952)

The general transistor properties of small size and weight, low power and voltage, and potential long life suggest extensive application of transistors to pulse or switching type systems of computer or computer-like nature. It is possible to devise simple regenerative circuits which perform the normally employed functions of waveform generation, level restoration, delay, storage (registry or memory), and counting. The discussion is limited to point contact type transistors in which the alpha or current gain is in excess of unity and to a particular feedback configuration. Such circuits, which are of the so-called trigger type, are postulated to involve negative resistance. On this basis an analysis, which approximates the negative resistance characteristic by three intersecting broken lines, is developed. Conclusions which are useful to circuit and device design are reached. The analysis is deemed sufficiently accurate for the first order equilibrium calculations. Transistors having properties specifically intended for pulse service in the circuits described have been developed. Their properties, and limitations, and parameter characterizations are discussed at some length.

INTRODUCTION

It is proposed to discuss some of the properties of transistors which are applicable to switching or pulse-type circuits, to develop elementary analysis methods and to describe a few circuits.

The bounds or limits of the field of switching are difficult to define. The common thread usually involves definite states of being as "open or closed", "off or on", "0 or 1", and so on, rather than a continuum of conditions. Even when consideration is given to more than two states, the thought involves distinct recognition of each state. The field is termed to be non-linear in distinction to linear manipulation of information. Any number of anomalies in definition may be raised.

Without attempting either to define or to limit the field, some of the functions which are often employed are: wave form generation, as rectangular pulses, sawtooth waves, etc.; memory or storage which may

be for short, intermediate or long periods and involves the retention of information for subsequent use; operations involving addition, subtraction, multiplication and division; translation of information from one form or code to another; gating, involving the routing of signals according to a predetermined pattern or set of conditions; regeneration of signals in amplitude and wave form; delay, which may be thought of as a form of storage; and timing. Some of these functions are simple; others result from fairly complex structures of simpler functions.

Present trends in electronic switching systems are toward complicated automata as exemplified by digital computers.¹ The reliability, power consumption and physical size of the electron devices employed largely determines the degree of realizability of such systems. It is believed that the transistor will find a significant application in this field.

The transistor can reduce power consumption by the elimination of heater or filament power. In addition, particularly in broadband applications as in high speed pulse systems, the "B" power may be reduced by the order of one or two decades if not more. Transistor circuits with $0.02\ \mu\text{s}$ rise time have been made to operate with an input power of 20 milliwatts which compares with approximately 2.5 watts (1-watt heater, 1.5-watt plate) for an equivalent tube circuit. Transistors have operated with less than one microwatt input power.²

Such power reductions result from the low operating voltages, low internal resistances and low capacitances of transistors. Low internal impedances greatly reduce the importance of stray wiring capacitances thereby making mechanical design much simpler and often eliminating the need for isolating or buffer amplifiers.

The transistor can contribute definite reduction in size directly. Fig. 1 shows a "bead" transistor which has a volume of approximately 1/1000 of a cubic inch and a weight of 5/1000 ounce. Indirectly the transistor can contribute to size reduction through the use of smaller, lower voltage, lower dissipation components. The reduction of power supply requirements in terms of size, regulation and capacity is also quite appreciable.

Transistors have been subjected to shocks in excess of 20,000 G without change in characteristics. Vibration tests have shown no resonances in the transistor shown in Fig. 1 to several thousand cycles. Harmonic accelerations of $100\ \text{G}$ at 1000 cycles have produced no detectable current modulation.

¹ L. N. Ridenour, "High Speed Digital Computers", *J. Appl. Phys.*, **21**, pp. 263-270, April, 1950.

² R. L. Wallace, Jr. and W. J. Pietenpol, "Some Circuit Properties and Applications of *n-p-n* Transistors", *Bell System Tech. J.*, **30**, pp. 530-563, July, 1951.

Life reliability and expectancy are difficult to determine due to the relative infancy of transistors, the definite finiteness of time, the many variables involved and the rate of development progress. Average life is presently estimated to be in excess of 70,000 hours. Life is a function of the operating conditions and may be materially reduced accordingly.

Transistors also have limitations. Noise at present is high for point-contact types as compared to electron tubes; input impedances are low, which may be either advantageous or disadvantageous; power output may be limited; frequency response is relatively low; circuit instability may cause design difficulties; and the devices are sensitive to temperature changes. There is also an absence of a long practical experience with a consequent art background in both devices and circuits.

A comprehensive review of transistor properties is given in the paper by J. A. Morton.³

While it is difficult to define the switching field, it is no less difficult to discuss circuit and device properties on a general basis. This is related to the non-linear nature of the circuits and devices in distinction to the virtually classical linear small-signal field. The lack of a classical method of analysis is a serious handicap in the synthesis of contemporary circuits and devices. When new devices, as the transistor, are to be considered,

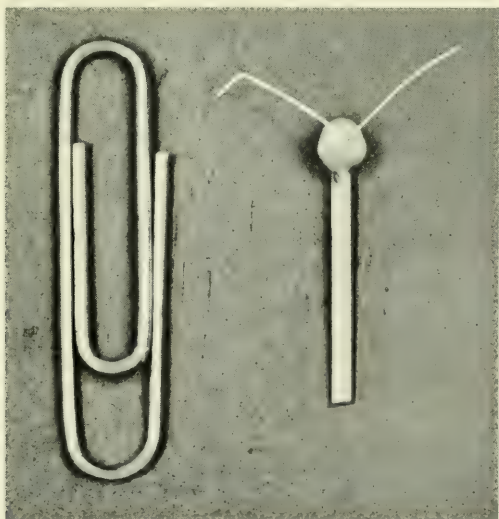


Fig. 1—A miniature switching-type transistor (M1689).

³ J. A. Morton, "Present Status of Transistor Development", *Bell System Tech. J.*, **31**, pp. 411-442, May, 1952.

the problem is multiplied due to the lack of a long background of experience.

It has been assumed that negative resistance is a common thread among "trigger circuits" and oscillators regardless of the device employed — electron tube, gas tube, transistor, mechanical structures, etc. This is not a new or novel idea and there is no intent to present it as such.⁴ Rather, it is used as a pattern upon which a certain degree of transistor analysis may be based, leading to simple understanding. The analysis assumes that the negative resistance characteristic can be broken into three regions; each region is then considered on a linear basis.

Section I will deal with simple circuit properties; Section II with analysis and Section III with device properties.

I—SIMPLE CIRCUIT PROPERTIES

The common property ascribable to switching functions is that of definiteness of state. The condition of the function is either "off" or "on". Switches are either open or closed; relays are operated or not; tubes are in cutoff or overload; doors are open or closed and so on. This is common regardless of the phenomena being exploited.

There is an intermediate region between these two conditions usually characterized by a time which is related to how fast the function may go from one state to another. Functionally the times of closing and opening are taken to be zero; practically, they are of determining importance. Relays replace hand-operated switches and electronic devices replace relays as speed becomes important. Obviously, no function or system can be faster than its state-devices.

All such state-devices will have separate attendant properties such as the degree of reverse coupling between the controlling signal and the controlled signal. Separated into families, however, there are those which are passive and those which are active. The latter are threshold devices in which a small amount of signal or control energy causes the translation of a relatively larger amount of stored energy into dynamic energy which consummates the change in state. As long as the control

⁴ See for example "Negative Resistances, Their Characteristics and Effects. Sinusoids, Relaxation Oscillations and Relaxation Discontinuities", Walter Reichardt, *Elektrische Nachrichten-Technik*, **20**, pp. 76–87, March, 1943; "Uniform Relationship Between Sinusoids, Relaxation Vibrations and Discontinuities", Walter Reichardt; *Elektrische Nachrichten-Technik*, **20**, pp. 213–225, Sept., 1943. For transistors: "Counter Circuits Using Transistors", E. Eberhard, R. O. Endres and R. P. Moore, *RCA Review*, pp. 459–476, Dec. 1949; "A Transistor Trigger Circuit", H. J. Reich and Ungvary, *Rev. Sci. Instr.*, **20**, p. 8, p. 586, Aug., 1949; and "Some Transistor Trigger Circuits", *Proc. Inst. Radio Engrs.*, **39**, pp. 627–632, June, 1951, P. M. Schultheiss and H. J. Reich.

signal is below the initial threshold there is no response and any change is directly related to the passive transmission of the control signal alone. When the signal exceeds the threshold the second state is assumed. Watch escapements, thyratrons, and the whole family of oscillators fall into this category. When the simplest cases of such functions are analyzed, they are found to involve in one way or another two stable states separated by a region in which there is positive feedback and gain in excess of unity with a resultant equivalent negative resistance. The proposition that a negative resistance characteristic is common to trigger or threshold switching circuits is tacitly assumed. The next step is to examine the transistor for such behavior and to classify the properties.

NEGATIVE RESISTANCE IN THE TRANSISTOR

That the transistor* can exhibit negative resistance has been demonstrated analytically⁵ and experimentally. The resistances seen looking into the emitter and collector of the transistor with grounded base are shown in Fig. 2.

In the equations and discussion to follow, the symbol conventions are as follows: External circuit elements are capitalized as R_e , R_b , and R_c . The symbols R_{11} , R_{12} , R_{22} and R_{21} define the open-circuit transistor resistances; the symbols r_e , r_c , r_m , and r_b define the equivalent circuit

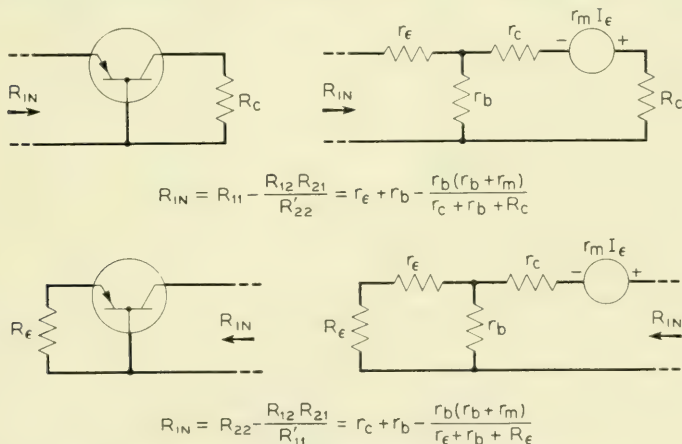


Fig. 2—Emitter and collector driving point resistances.

* Discussion is limited primarily to point contact transistors with α 's or current gains greater than unity.

⁵ R. M. Ryder and R. J. Kircher, "Some Circuit Aspects of the Transistor", *Bell System Tech. J.*, **28**, pp. 367-400, July, 1949.

element values. Network resistances which contain both device and external elements are primed. For example, $R'_{22} = R_{22} + R_c + R_b$, where $R_{22} = r_c + r_b$. See also references 3 and 5.

Taking the collector or output resistance, Fig. 2, for example,

$$R_{in} = (r_c + r_b) - \frac{r_b(r_b + r_m)}{r_e + r_b + R_e} \quad (1)$$

R_{in} can be negative or positive depending upon the relative magnitudes of the two terms. Actually, of course, r_m has a phase factor and so is frequency dependent. Frequencies wherein r_m is essentially resistive will be assumed. For negative resistance, r_m must be large, R_e small and r_b not too small or else augmented by external resistance. Negative resistance is thus predicted on a small-signal linear basis. The large-signal behavior may be studied experimentally by adding sufficient resistance as R_c to the first or positive term to insure stability. This is shown in Fig. 3 with the resultant characteristic. External base resistance R_b has been added and R_e is zero.

Fig. 3 illustrates the pattern of a three-valued characteristic: Regions I and III are portions with positive slope, indicating stable operating

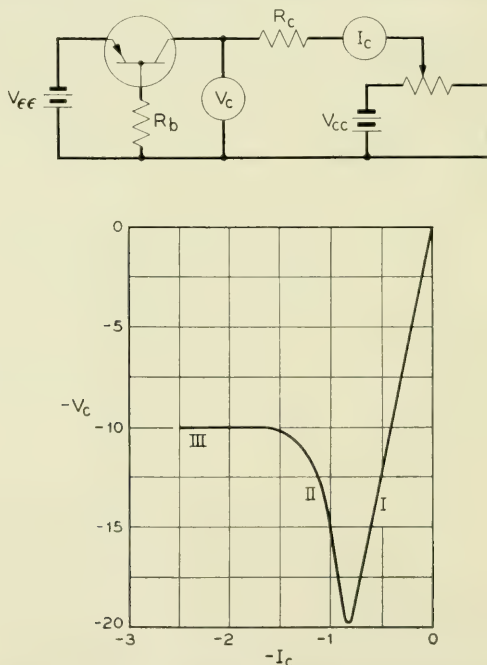


Fig. 3—Collector large-signal negative resistance characteristic.

regions, separated by Region II, a region of negative slope, indicating the possibility of instability. In this particular case, Region I has high resistance and Region III very low resistance.

An evaluation of the emitter or input characteristic leads to similar results, using the circuit of Fig. 4. R_b has been added here also and R_c taken as zero. The general pattern is again present. Region I has high, positive resistance; Region II, negative resistance; and Region III very low, positive resistance.

BIASES AND LOAD LINES—BISTABLE OPERATION

The negative resistances of Figs. 3 and 4 are both of the so-called open-circuit stable type. If loads are applied to the circuit terminals of Fig. 2 which are larger in magnitude than the negative resistances, the circuits will be stable; that is, there will be single operating points. This is shown in Fig. 4 by the dashed load lines marked, R'_ϵ , R''_ϵ , R'''_ϵ . A load resistance smaller in magnitude than the negative resistance may intersect the characteristic in three positions as shown by the load line R_ϵ .

The load line R_ϵ can be made to have single or multiple intersections by biasing properly as shown in Fig. 5, where the three possibilities are shown as R_ϵ , R'_ϵ , R''_ϵ . Single or multiple intersections result in accordance with the choice of emitter bias, V_{ee} , as shown. It can be shown

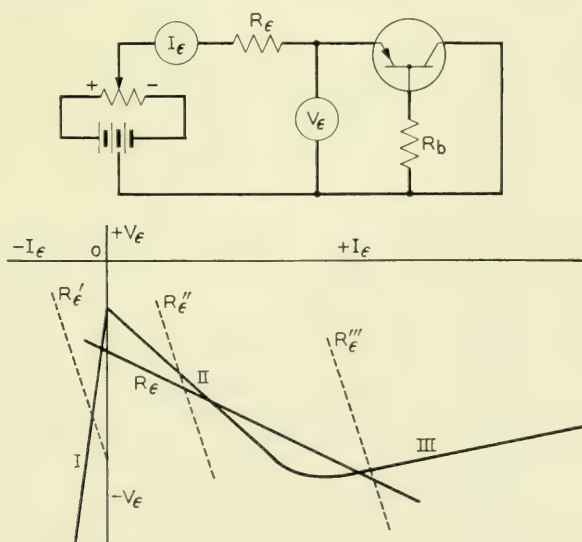


Fig. 4—Idealized emitter large-signal negative resistance characteristic.

that the intersection of load line R_ϵ with the characteristic at b in Fig. 5 is unstable whereas those at a and c are stable. Experiment in the multiple intersection case shows also that as $V_{\epsilon\epsilon}$ is slowly increased (decreased in absolute magnitude) the load line moves upward and that the assumed operating point, a , moves up along the Region I portion of the characteristic. At the turning point shown on the current axis, the operating point suddenly flips to the high current region, returning along the curve to c as $V_{\epsilon\epsilon}$ is returned to the original value.

A decrease in $V_{\epsilon\epsilon}$ toward $V_{\epsilon\epsilon}''$ moves the operating point at c downward along the characteristic until it "escapes" past the lower turning point and flips to the Region I portion, returning to a as $V_{\epsilon\epsilon}''$ is returned to the original value. This then is an elementary switching circuit, a bistable trigger circuit or "flip-flop". A positive emitter pulse will cause the circuit to flip to high current, a negative pulse to low current. The triggers may be applied to emitter, base or in combination with proper attention to polarity. Trigger sensitivities are shown in Fig. 6. Such a circuit is often used for register or storage purposes. It can store one bit of information for a potentially infinite period, be sampled for the presence of such information, and be cleared or restored to the original condition for reuse when the stored information is no longer useful.

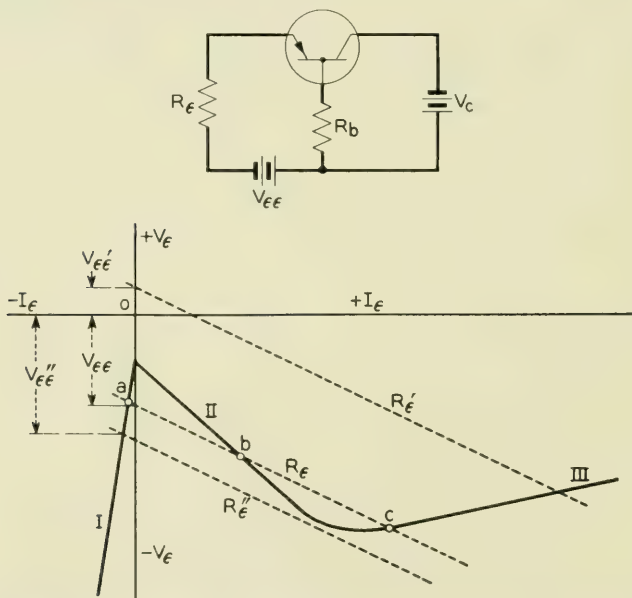


Fig. 5—Emitter negative resistance characteristic showing possible multiple operating points.

With the addition of suitable steering circuits it can be made to count by a scale of two.

MONOSTABLE AND ASTABLE CIRCUITS

The addition of a capacitor to the circuit as in Fig. 7(a) leads to either monostable or astable operation. In Fig. 7(b) the normal operating point is stable at a as discussed previously by virtue of the bias V_{EE} . As V_{EE} is increased, as by a trigger, the load line is moved up and over the turning point. Without capacitor C in the circuit, the operating point would move to b with the resultant rapid change in voltage and current. However, a capacitor has in effect voltage inertia; this is equivalent to saying that a capacitor is a short-circuit to a voltage change. Both the capacitance and the rate of change of voltage are assumed high. Thus at the turning point the capacitor effectively short-circuits the emitter and the operating point snaps along dotted line (1) to intersect the characteristic. This point is quasi-stable and the capacitor is discharged along line (2) to the second turning point where the emitter is again effectively short-circuited and the operating point snaps along (3) to intersect the Region I portion of the characteristic. This point is also quasi-stable and the operating point moves slowly up to the initial or dc stable operating point. A single trigger thus causes a complete cycle of operation. The emitter current shifts

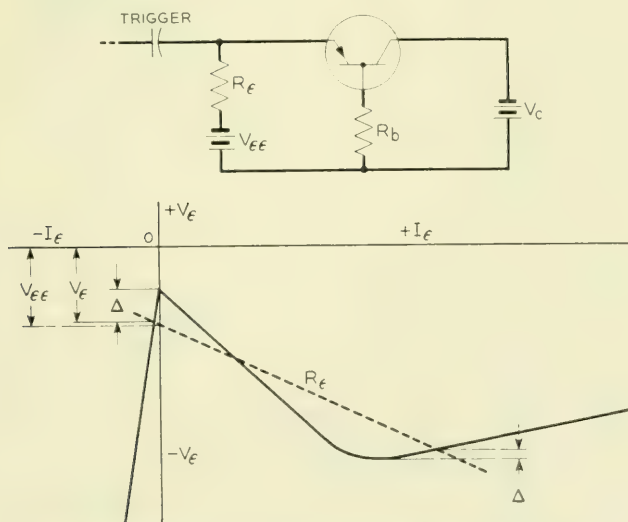


Fig. 6—Bistable circuit showing trigger sensitivity, Δ .

rapidly to a high value of current, falls relatively slowly to an intermediate value, then shifts rapidly to a small negative value and finally returns slowly to the original value. The emitter current and voltage are sketched in Fig. 8. It is a so-called "single-shot" circuit. Alternately the rest or dc stable point can be chosen to be in Region III, at high current, by choice of positive instead of negative bias V_{ee} . Practical considerations as ease in triggering and average power consumption usually indicate a preference for the Region I dc stable point.

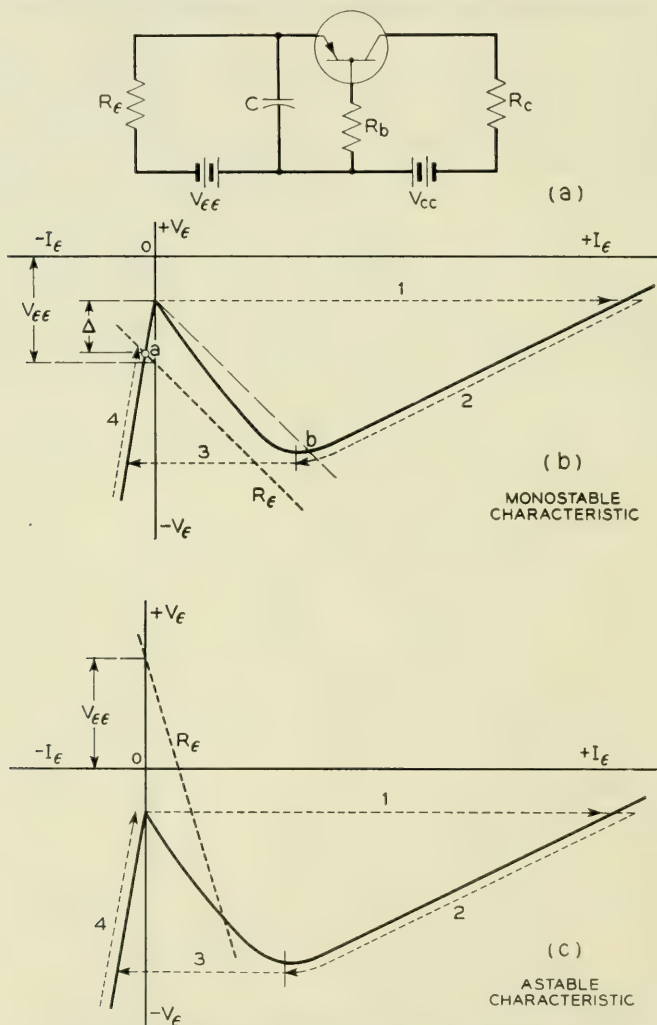


Fig. 7—Monostable and astable characteristics.

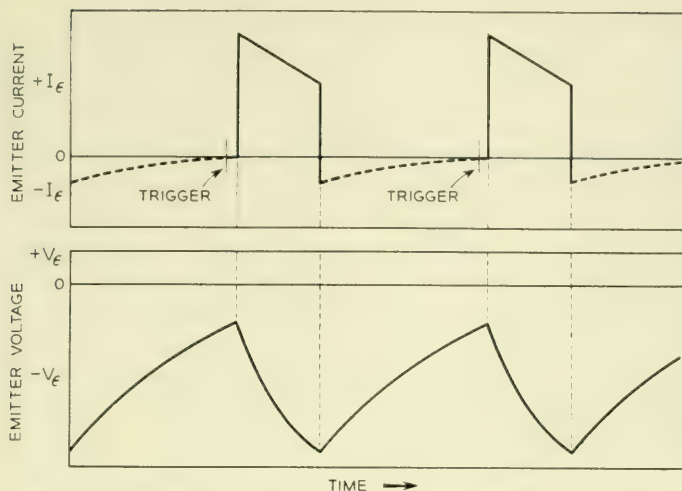


Fig. 8—Idealized monostable relaxation oscillator waveforms.

When the load line and bias are chosen to result in intersection in the negative resistance portion, astable operation or continuous oscillation results. This mode is illustrated in Fig. 7(c). Proper bias and $R_e > |-R_{in}|$, Region II, are required. The operating point formed by the intersection of the load line on the negative resistance portion of the characteristic would normally be stable. However, the capacitor provides an ac short-circuit in parallel with R_e causing the path (1), (2), (3), (4) to be followed continuously. Another form of physical explanation of this relaxation oscillation, usually applied to gas tubes, is that the capacitor C is charged slowly through R_e to a critical or breakdown value whereupon the tube or device rapidly discharges the capacitor. When the capacitor charge is dissipated, the device discharge can no longer be maintained due to the IR drop in R_e and the tube or device open-circuits and the capacitor is recharged.

The above suggests a strong similarity to gas tube behavior and this is indeed so. In fact, the modes described above are common to all open-circuit stable negative resistance devices; only the parameters and device phenomena are different.

The primitive circuits of Fig. 7 have properties basic to several switching functions. These may be deduced from the waveforms of Fig. 8 which are essentially identical to both the monostable and astable cases. The emitter current has a rectangular waveform which suggests the generation of rectangular pulses; and, for the astable case, regenerative amplification for both amplitude and wave shape, pulse rate or

frequency division and delay. As shown the current waveform is not particularly good, having neither a flat top nor a flat base line. Practically, the waveform may be derived from the collector by means of a small load resistor to obtain a flat base line. When the emitter current is negative there is sensibly no transfer action, hence, the collector current will be constant during the re-charge portion of the cycle instead of exponential as shown. The slope of the top is inherent and may be removed by clipping. Pulse rise time, the time required for transition from low current to high current, of $0.1 \mu\text{s}$ is quite easily obtainable; $0.02 \mu\text{s}$ with average input powers of 20 mw have been obtained. Fall time is usually longer than the rise time by factors of 3 or 4. It is to be noted in Fig. 8 that there has been shown a delay between the trigger application and the current transition. Such delay is not peculiar to transistors, but is common to all trigger type devices and circuits. The delay is shown here exaggerated in order to establish its existence and is associated with the static charging of the circuit and the dynamic delay of the device concerned. The trigger-transition delay with transistors is usually less than $0.1 \mu\text{s}$.

The voltage waveform of Fig. 8 has a sawtooth form and may thus be employed to generate linear time bases or sweeps. The normal methods for linearization such as a high charging voltage V_{cc} and a high charging resistance R_c or other constant current means are applicable here as in other device circuitry. Free-running and driven sweeps may be obtained with the astable and monostable circuits respectively.

Since the collector characteristic shown in Fig. 3 is also open-circuit stable, the same sort of circuits can be constructed using the output characteristic. Bistable, monostable and astable circuits are shown in Fig. 9.

The resistances seen looking into the base are given in Fig. 10. These circuits are short-circuit stable. That is, high values of R_b result in instability. Bistable, monostable and astable circuits can be constructed also, but use is made of an inductor instead of a capacitor. The reactance of the inductor affords a quasi-open-circuit in the same manner as the capacitor afforded a quasi-short-circuit in the previous cases. Circuit examples are shown in Fig. 11.

SUMMARY

These simple circuits by no means exhaust the switching circuit possibilities of the transistor; rather, they are the simplest. The simple circuit is often satisfactory and may sometimes be employed with little more understanding than that given. More often, however, problems

relating to the sensitivity, constancy of sensitivity, operating currents and voltages, interchangeability and the like require a much more quantitative understanding in order to create circuit designs having specific properties.* An equal need also exists in transistor design for analytic circuit relationships. Such information is useful first, in the

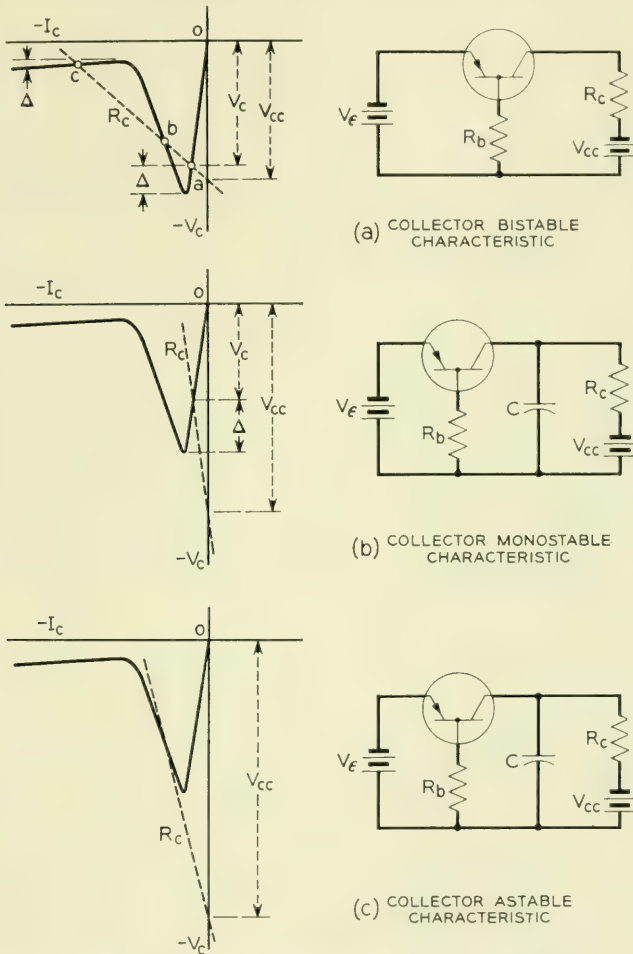


Fig. 9—Collector connection switching circuits.

* See, for example, J. R. Harris, "A Transistor Shift Register and Serial Adder", *Proc. IRE*, Nov., 1952; R. L. Trent, "A Transistor Reversible Binary Counter", *Proc. Inst. Radio Engr.*, Nov., 1952; H. G. Follingstad, J. N. Shive, R. E. Yaeger, "An Optical Position Encoder and Data Transmitter", *Proc. Inst. Radio Engr.*, Nov., 1952.

creation of optimized designs and, second, in the maintainance of proper parameter controls in manufacture. Finally, the more detailed the understanding, the more likely will be the creation of new circuits and new devices.

A complete analytical treatment will not be attempted here; consideration will be limited to the equilibrium case and in particular to the simple circuits described.

II—ANALYSIS

In order to deal analytically with circuits and devices it is necessary to have analytic expressions for the device characteristics. For small signal analysis this is relatively easy. In large signal applications, as in switching, the situation is not so simple. The problems arise because of the high degree of nonlinearity wherein the simplifying assumptions employed in small signal analysis are by no means valid. Further, it is desirable to retain dc terms in many cases.

The method to be employed here is the so called broken-line method which involves approximating the negative resistance characteristic by three intersecting straight lines. The assumption is made that there are three distinct regions of operation in each of which the device is separately linear, but involving different parameter values for each region.

The approximation is shown in Fig. 12. The assumption that the negative resistance characteristic can be simulated by three straight lines is reasonably valid for gross considerations; for fine detail near the

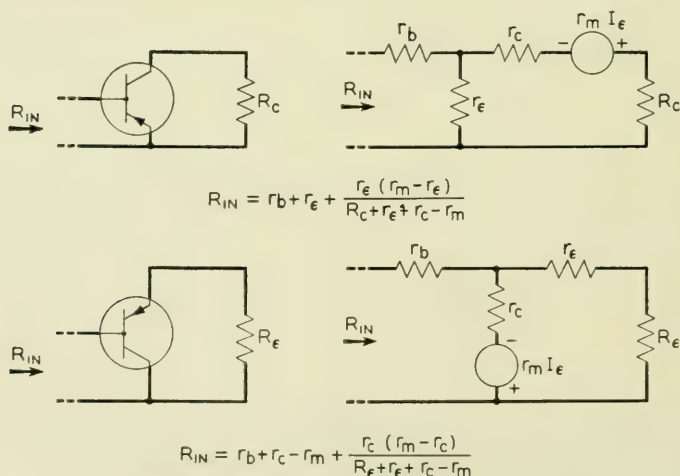
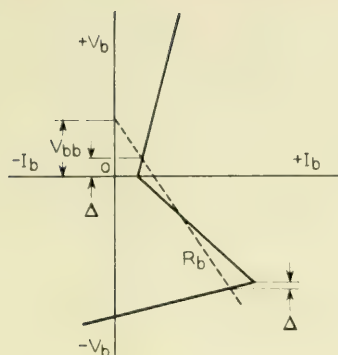
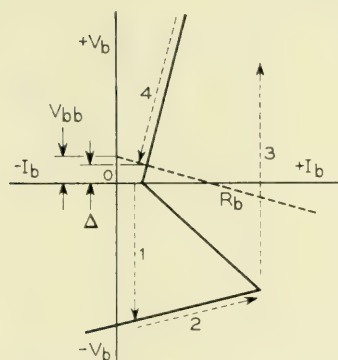
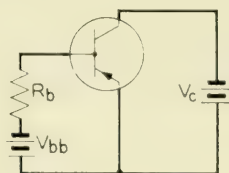


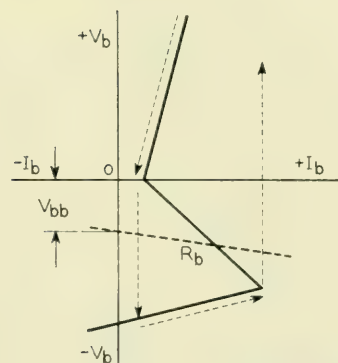
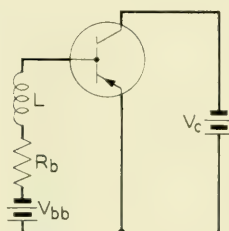
Fig. 10—Base driving point resistances.



(a) BASE BISTABLE CHARACTERISTIC



(b) BASE MONOSTABLE CHARACTERISTIC



(c) BASE ASTABLE CHARACTERISTIC

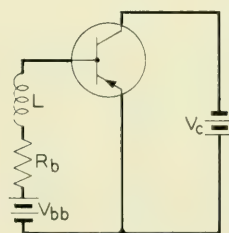


Fig. 11—Base connection switching circuits.

turning points the approximation is by no means accurate although affording zero order information.

Preparatory to the analysis of the negative resistance characteristics, it is necessary to obtain analytic expressions for the transistor currents and voltages. This in turn involves the following steps:

1. Identification of the three regions in terms of the device characteristics,
2. Idealization of the device characteristics to obtain simple, linear relations, and
3. Evaluation of the device parameters in each of the three regions.

Fig. 13 is a family of open circuit characteristics for a typical switching type transistor. Specifically, in small signal terms,

TABLE I

<i>Parameter</i>	<i>Equivalent Tee</i>
$R_{11} = \left. \frac{\partial V_e}{\partial I_e} \right]_{I_e}$	$R_{11} = r_e + r_b$
$R_{12} = \left. \frac{\partial V_e}{\partial I_c} \right]_{I_e}$	$R_{12} = r_b$
$R_{21} = \left. \frac{\partial V_c}{\partial I_e} \right]_{I_e}$	$R_{21} = r_m + r_b$
$R_{22} = \left. \frac{\partial V_c}{\partial I_c} \right]_{I_e}$	$R_{22} = r_c + r_b$

Also
$$\alpha = - \left. \frac{\partial I_c}{\partial I_e} \right]_{V_e} = \frac{R_{21}}{R_{22}} = \frac{r_m + r_b}{r_c + r_b}$$

The above set, normally employed for small signal analysis, will be assumed to be constant within a given region, but changing in value from region to region.

IDENTIFICATION OF THE THREE REGIONS

It may be recalled with the aid of Fig. 12 that the negative resistance characteristic consists of a negative resistance region bounded on each side by a region of positive resistance. Thus the device is first passive in nature with little or no gain, then very abruptly exhibits considerable gain with the resultant negative resistance, and finally becomes very abruptly passive again with little or no gain.

It would seem quite clear that abrupt changes in the transmission properties of a device should be associated with equally abrupt changes

in the forward transfer characteristic. In the case of the transistor, the behavior of the forward transfer properties is given by the forward transfer impedance, R_{21} .

Examining the R_{21} family in Fig. 13, it is seen that in the normal, positive emitter current region the slope, R_{21} , is high indicating the possibility of high forward gain. When I_e is negative, however, the slope is zero or nearly so, changing very abruptly at $I_e = 0$. Further, it is to be noted that as I_e is made negative, the collector voltage is unaffected, remaining constant for further change in I_e . Thus it may be said that the collector voltage is saturated.*

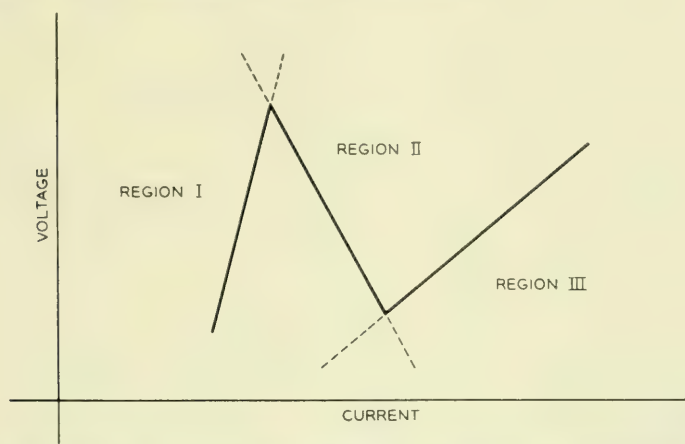


Fig. 12—Broken-line idealization of negative resistance characteristic—division into regions.

If, on the other hand, the emitter current is increased, at constant collector current, it is found that at a critical emitter current the slope again becomes zero or nearly so. There are also two further observations. First, the collector voltage is reduced to a very small value and second, that the critical emitter current is related to the collector current. From the small-signal relation,

$$V_c = R_{21}I_e + R_{22}I_c \quad (2)$$

or

$$V_c = R_{22}\alpha I_e + R_{22}I_c, \quad (3)$$

* It is tacitly assumed that in the relation $y = f(x)$ that there are extremes at which y becomes essentially constant and independent of further change in the independent variable x . The point farthest removed from the origin at which the dependent variable becomes constant is termed saturation. The point closest to the origin at which the dependent variable becomes constant is termed cutoff.

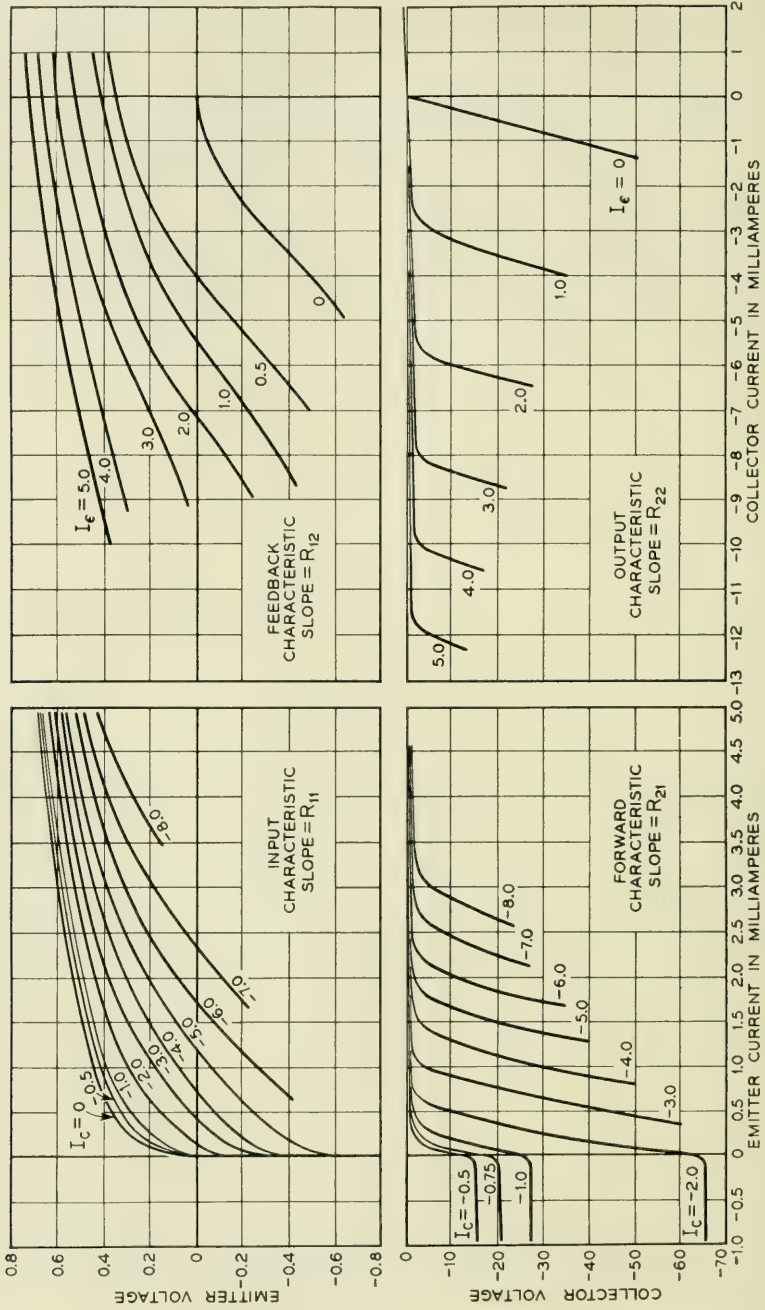


Fig. 13—Static characteristics of the MI689 developmental switching type transistor.

the critical emitter current for collector voltage cutoff may be obtained by setting $V_c = 0$, as,

$$I_\epsilon = -\frac{I_c}{\alpha} \quad (4)$$

This relationship is dual to the grid voltage-plate voltage relation in tubes for plate current cutoff as, $V_g = -(V_p/\mu)$. The criteria for defining the three regions are thus established as:

$$\text{Region I (Collector Voltage Saturation): } I_\epsilon < 0 \quad (5)$$

$$\text{Region II (Active): } 0 < I_\epsilon < -\frac{I_c}{\alpha} \quad (6)$$

$$\text{Region III (Collector Voltage Cutoff): } I_\epsilon > -\frac{I_c}{\alpha} \quad (7)$$

The identification of device parameters will be made for the several regions by a single prime for Region I as r'_ϵ , none for Region II as r_ϵ , and three primes for Region III as r'''_ϵ .

LINEARIZATION OF THE CHARACTERISTICS AND APPROXIMATIONS

The next step is to linearize the characteristics and to make suitable approximations in order to obtain simple linear equations of the terminal currents and voltages. The relations which require linearization are the device parameters R_{11} , R_{12} , R_{21} and R_{22} which are in general functions of the currents as $R_{ij} = f(I_1, I_2)$.

LINEARIZATION OF R_{11} AND R_{12}

In terms of the equivalent tee circuit, which has been and will be employed, R_{11} is given as $R_{11} = r_\epsilon + r_b$. Also, $R_{12} = r_b$. It is convenient to separate r_ϵ and r_b and discuss each separately since r_b is fairly constant and r_ϵ will have widely different regional values.

In the R_{12} family of Fig. 13, it may be seen that R_{12} or r_b is fairly constant in all three regions and will be so taken here. Further, in the simple circuits under consideration, external base resistance R_b has been inserted so that minor variations in r_b in the total of $r_b + R_b$ are inconsequential since usually $R_b \gg r_b$. The approximation that r_b is constant is subject to review where finer detail is necessary, particularly at low emitter currents where the rate of change of r_b is at a maximum.

The emitter resistance r_ϵ is approximately the resistance of a diode in the forward direction. As such, r_ϵ is high when the emitter current is

negative and low when the emitter current is positive. Experimentally, it is found convenient to give three values to r_e and hence to R_{11} , one for each region as shown in Fig. 14. This recognizes the non-linearity with I_e in the forward direction and assumes that a single value in the reverse direction is sufficient. As the circuitry becomes more sophisticated a more precise approximation will undoubtedly be required, particularly near $I_e = 0$.

It may be noted that in the functional relation $R_{11} = f(I_e, I_c)$ that R_{11} is taken to be a function of I_e only. The contribution of I_c is to shift the characteristic in voltage by $r_b \Delta I_c$ increments. Thus the relationship of $V_e = f(I_e, I_c)$ can be written very simply as

$$V_e = R_{11}I_e + R_{12}I_c \quad (8)$$

Since the problem has been linearized to first order terms only, the currents and voltages are total instantaneous or dc values as indicated by the capital letters.

IDEALIZATION OF R_{21}

As indicated previously, R_{21} will be small in Regions I and III and large in Region II. Since $R_{21} = r_m + r_b$, R_{21} can be no less than r_b ; the defining approximations will be applied to r_m . In Region I when the emitter current is negative, r_m is taken to be zero and reflects the device approximation that the emitter current under this condition is entirely electron current. This is not always a true approximation, particularly near $I_e = 0$, and limiting tests are employed in transistor testing.

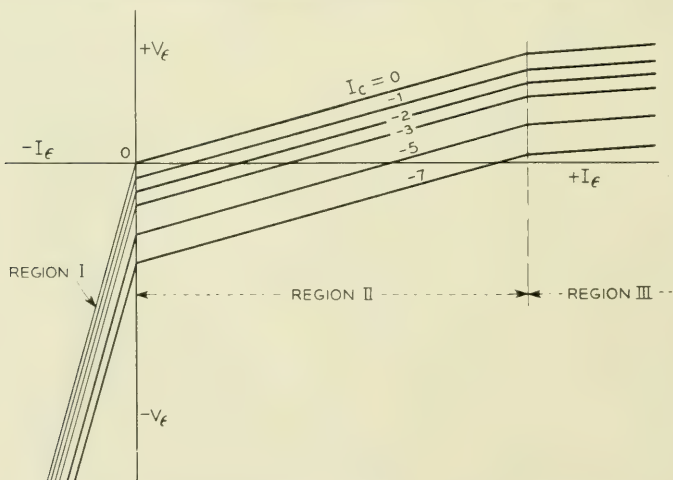


Fig. 14—Idealization and regional division of input characteristic (R_{11}).

In Region II, r_m has, of course, high values and in general $r_m \gg r_b$. Pending further investigation, r_m will be assumed finite, but very small in Region III.

IDEALIZATION OF R_{22}

In the output family of Fig. 13 it may be noted that R_{22} has two values, a high value for $I_c > -\alpha I_e$ and a low value for $I_c < -\alpha I_e$. The high value corresponds to Regions I and II and the low value to Region III. To a first order the two values are separately constant which was not true of earlier transistors in which R_{22} underwent extensive degradation in magnitude as I_c and I_e increased.

The lower limit to which R_{22} can fall in Region III is r_b , since $R_{22} = r_c + r_b$, implying that r_c is zero in Region III. This is approximately, but not accurately true. As αI_e approaches $-I_c$ in magnitude, the voltage across the collector barrier becomes nearly zero so that r_c has a low, but finite value. Under this condition, the hole current is very high and heavy conductivity modulation of the collector barrier resistance occurs. Thus the collector resistance in Region III is indeed quite low and may be neglected for many circuit computations.

In the functional relation $R_{22} = f(I_e, I_c)$ it has been assumed that R_{22} is a function of I_c alone. Further, the approximation involves first order terms only and hence the functional relation $V_c = f(I_e, I_c)$ may be written as:

$$V_c = R_{21}I_e + R_{22}I_c \quad (9)$$

Here again, as in the input case, the currents and voltages are total instantaneous or dc values as indicated by the capital letters.

It is believed desirable, however, to give one more consideration to the output relations. When $I_e = 0$, the collector characteristic is approximately that of a diode in the reverse direction. A diode has low reverse resistance until the voltage across the barrier exceeds a few tenths of a volt and then has quite high resistance, approaching infinite slope in the case of junction diodes.^{6, 7} This effect is shown exaggerated in the idealized output family of Fig. 15. The current and voltage at the break in the $I_e = 0$ curve have been termed I_{c0} and V_{c0} respectively. I_{c0} and V_{c0} are quite evident in junction devices; in point contact devices they are not nearly so evident due to the lower value of R_{22} and the higher voltages and currents normally employed. Where currents and

⁶ See Reference 2.

⁷ *Holes and Electrons*, W. Shockley, Van Nostrand, p. 91, 1950.

voltages are of the order of several milliamperes and a few volts, I_{c0} and V_{c0} may normally be neglected. I_{c0} and V_{c0} do have considerable interest to the device designer, however. The net circuit interpretation of I_{c0} and V_{c0} is to effectively transfer the current-voltage axis from 0, 0 to I_{c0} , V_{c0} . Therefore,

$$V_c - V_{c0} = R_{21}I_\epsilon + (I_c - I_{c0})R_{22} \quad (10)$$

or

$$V_c - V_{c0} = (r_m + r_b)I_\epsilon + (I_c - I_{c0})(r_c + r_b) \quad (11)$$

Making the approximation that $V_{c0} = I_{c0}R_{22}'$ and rearranging, equation (10) becomes,

$$V_c - I_{c0}R_{22}'' + I_{c0}R_{22} = R_{21}I_\epsilon + I_cR_{22} \quad (12)$$

or

$$V_c + I_{c0}(R_{22} - R_{22}'') = R_{21}I_\epsilon + I_cR_{22} \quad (13)$$

which is of the usual form except that a small dc generator of magnitude $I_{c0}(R_{22} - R_{22}'')$ has been added in series with the collector. Since $R_{22} = r_c + r_b$ and $R_{22}'' = r_c'' + r_b$,

$$I_{c0}(R_{22} - R_{22}'') = I_{c0}(r_c - r_c'') \quad (14)$$

The output family equation with equivalent circuit parameters is

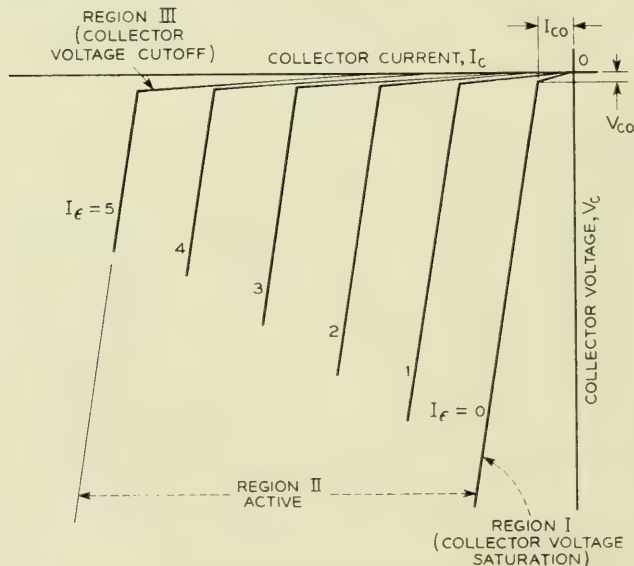


Fig. 15—Idealization and regional division of output characteristic (R_{22}).

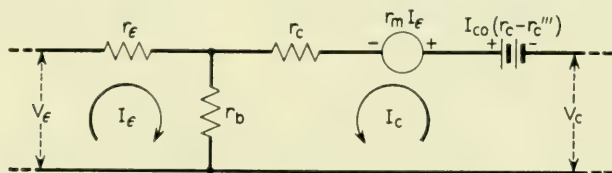
then:

$$V_c + I_{c0}(r_c - r_c''') = (r_m + r_b)I_\epsilon + (r_c + r_b)I_c \quad (15)$$

SUMMARY OF IDEALIZATION OF CHARACTERISTICS

The results of the idealization of the device characteristics are summarized in Fig. 16. Here are given analytic expressions for the input and output voltages in terms of the input and output currents; the regions are defined symbolically and by typical values; and an equivalent circuit is given. It may be noted that the equivalent circuit is identical to the small-signal equivalent tee, excepting the small dc generator $I_{c0}(r_c - r_c''')$ which usually may be neglected when dealing with contemporary point contact transistors.

To obtain any of the negative resistance characteristics it is only necessary first to solve the two equations simultaneously for the appropriate voltage in terms of the appropriate current, and then second to insert into the resultant equation the proper parameter values, region by region, to obtain three equations. These equations, when plotted, result in an idealized characteristic similar in form to that of Fig. 12. A detailed example plus synopsis of the properties of the several connections will be given in the following sub-section.



$$V_\epsilon = (r_\epsilon + r_b) I_\epsilon + r_b I_c$$

$$V_c + (r_c - r_c''') I_{c0} = (r_m + r_b) I_\epsilon + (r_c + r_b) I_c$$

REGION	PARAMETER							
	r_ϵ		r_b		r_c		r_m	
	SYMBOL	TYPICAL	SYMBOL	TYPICAL	SYMBOL	TYPICAL	SYMBOL	TYPICAL
I	r_ϵ'	100 K	r_b	160	r_c	20 K	r_m'	0
II	r_ϵ	100	r_b	160	r_c	20 K	r_m	50 K
III	r_ϵ'''	25	r_b	50	r_c'''	70	r_m'''	30

$$I_{c0} \approx -50 \mu A$$

Fig. 16--Broken-line transistor equations, regional parameter values and equivalent tee circuit.

THE ALPHA OR CURRENT GAIN FACTOR*

The derivation just given has been in terms of the equivalent circuit parameters, r_e , r_b , r_c , and r_m . Another circuit factor, alpha or the short-circuit current gain, is also quite useful. Alpha has been defined in Table 1 as the negative ratio of the incremental change in output current to the incremental change in input current *under the condition of short-circuit output terminals*.

Thus alpha is restricted in interpretation to a specific device termination and care should be taken in the employment of alpha when other terminations are involved. For example, the circuit current gain under general conditions is given by R'_{21}/R'_{22} . The ratio R'_{21}/R'_{22} has been sometimes termed α_c . Thus,

$$\alpha_c = \frac{R'_{21}}{R'_{22}} = \frac{r_b + R_b + r_m}{r_b + R_b + r_c + R_c} \quad (16)$$

Depending upon the magnitudes of R_b and R_c , the two current gain ratios may be markedly different. In Region II where r_m and r_c are very large the effects of R_b and R_c in equation (16) often may be neglected. The circuit current gain, α_c , may then be taken as the device alpha. In Region I, r_m has been taken as zero; hence the current gain will be somewhat less than unity, given by $(r_b + R_b)/(r_b + R_b + r_c + R_c)$, and is definitely not zero. Equally, in Region III, the circuit current gain is not zero but rather approaches the ratio, $(r_b + R_b)/(r_b + R_b + R_c)$. If $R_b \gg R_c$, the ratio is nearly unity.

ANALYSIS OF NEGATIVE RESISTANCE CHARACTERISTIC

The objectives of the circuit analysis, as stated previously, are:

1. To determine operating conditions such as proper values of loads, biases, trigger sensitivities and operating currents and voltages,
2. To determine the relationships of the device parameters to the circuit behavior in order that these parameters may be optimized, properly characterized and controlled for required circuit performance.

For example, the trigger sensitivity may be given by the voltage difference between the load line intersection with the Region I portion of the characteristic and the upper peak or turning point of the characteristic as shown in Figs. 6, 7 and 9. The sensitivity Δ is thus determined by the nearness of the bias point to the peak of the characteristic. Since the bias is normally fixed, variations in the sensitivity will arise

* This section is inserted parenthetically as clarifying material due to the use of the α -factor in subsequent discussion.

from variations in the peak point. Thus it is necessary to know the relationships which determine the currents and voltages of the peak and valley points in order first to achieve a design and second, to establish controls on the proper device parameters.

In this example the emitter negative resistance characteristic will be solved and analyzed. The solutions for the other characteristics follow in the same manner and will be summarized.

EVALUATION OF EMITTER CHARACTERISTIC AS AN EXAMPLE

To obtain the emitter characteristic, it is necessary to solve the two equations of Fig. 16 for V_e in terms of I_e . The equations as given are for the short-circuit case. Since the general case will involve external parameters as R_e and R_c , the equations will be modified to include these parameters.

The effects of external parameters may be applied very easily since,

$$V_e = V_{ee} - I_e R_e \quad (17)$$

and

$$V_c = V_{cc} - I_c R_c \quad (18)$$

where V_{ee} and V_{cc} are supply voltages; V_e and V_c are measured from the appropriate terminal to the far end of the external base resistor. External R_b adds directly to r_b . Thus the two equations become:

$$V_{ee} = (r_e + R_e + r_b + R_b)I_e + (r_b + R_b)I_c \quad (19)$$

$$V_{cc} + (r_c - r_e''')I_{c0} = (r_m + r_b + R_b)I_e + (r_b + R_b + r_c + R_c)I_c \quad (20)$$

In manipulation of equations (19) and (20) it is often more easy to do so in the functional form,

$$V_1 = R'_{11}I_1 + R'_{12}I_2 \quad (21)$$

$$V_2 = R'_{21}I_1 + R'_{22}I_2 \quad (22)$$

with substitution at the evaluation stage. The R'' 's here include both device and circuit parameters.*

Solving equations (19) and (20) simultaneously, the following rela-

* Here the primes indicate generalized open circuit driving point resistance rather than reference to Region I. The duplication of symbols is regretted.

tionship between V_ϵ and I_ϵ is obtained:

$$V_\epsilon = \left[r_\epsilon + R_\epsilon + r_b + R_b - \frac{(r_b + R_b)(r_b + R_b + r_m)}{r_b + R_b + r_c + R_c} \right] I_\epsilon + \frac{(V_{cc} + I_{c0}(r_c - r_c'''))}{r_b + R_b + r_c + R_c} (r_b + R_b) \quad (23)$$

Equation (23) is general for the given circuit; it suffers, however, in difficulty in interpretation due to the numerous terms. In the regional evaluation to follow, approximations will be made which bring out the significant factors although decreasing the accuracy somewhat. The $(r_c - r_c''')I_{c0}$ terms will be neglected. It is assumed also that large external base resistance R_b is employed.

EVALUATION IN REGION I

In Region I, from Fig. 16, r_m is zero and r'_ϵ is large so that $r'_\epsilon \gg (r_b + R_b)$. Also, by assumption, $r_b \ll R_b$. Applying these approximations, equation (23) becomes,

$$V_{\epsilon I} \approx r'_\epsilon I_\epsilon + \frac{V_{cc}R_b}{R_b + r_c + R_c} \quad (24)$$

Equation (24) is the equation of a straight line, having slope r'_ϵ and an intercept on the voltage axis at $(V_{cc}R_b)/(R_b + r_c + R_c)$. The small-signal input impedance is just the slope value or r'_ϵ .

The short-circuit case where R_c is zero is the most adverse device condition in the sense that the dc term will then be most dependent upon device parameters. When $R_c = 0$, equation (24) becomes

$$V_{\epsilon I} \approx r'_\epsilon I_\epsilon + \frac{V_{cc}R_b}{R_b + r_c} \quad (25)$$

EVALUATION IN REGION II

In Region II all parameters are finite and the only approximations which may be made are $r_b \ll R_b$ and $r_\epsilon \ll R_b$. Thus,

$$V_{\epsilon II} \approx \left[R_b - \frac{R_b(R_b + r_m)}{R_b + r_m} \right] I_\epsilon + \frac{V_{cc}R_b}{R_b + r_c + R_c} \quad (26)$$

If R_b is not too large, it may be approximated that $(R_b + r_m)/(R_b + r_c) = \alpha$. Taking $R_c = 0$, thus,

$$V_{\epsilon II} \approx R_b(1 - \alpha) + \frac{V_{cc}R_b}{R_b + r_c} \quad (27)$$

Equation (27) is also the equation of a straight line having the voltage axis intercept of $(V_c R_b)/(R_b + r_c)$ the same value as in Region I. The slope, $R_b(1 - \alpha)$, is negative provided $\alpha > 1$.

EVALUATION IN REGION III

In Region III it may be assumed that $r_b \ll R_b$, $r_c''' \ll R_b$ and $r_m''' \ll R_b$. Other suitable approximations will depend largely upon the magnitude of R_c . From equation (23)

$$V_{cIII} \approx \left[r_c''' + R_b - \frac{R_b(R_b + r_m''')}{R_b + r_c''' + R_c} \right] I_c + \frac{V_{cc} R_b}{R_b + R_c} \quad (28)$$

If R_c is large, that is, large compared to r_c''' , but small compared to Region II r_c , then (28) becomes,

$$V_{cIII} \approx \frac{R_b R_c}{R_b + R_c} I_c + \frac{V_{cc} R_b}{R_b + R_c} \quad (29)$$

Under these conditions, the circuit is essentially independent of device parameters. This is useful where a high independence of device parameters is required, but does not focus the attention upon the device parameters as does the $R_c \rightarrow 0$ case. This is the condition under which the transistor might be operated when it is desired to obtain the maximum ON current, or conversely the minimum internal switch resistance.

Where $R_c = 0$, equation (28) becomes,

$$V_{cIII} = [r_c''' + r_c''' - r_m'''] I_c + V_c \quad (30)$$

Since r_c''' and $(r_c''' - r_m''')$ are quite small the short-circuit currents may be very high. Where the transistor is considered as a switch between emitter and collector circuits, the "switch" voltage drop, as V_{ec} , is given by the first term of equation (30).

EVALUATION OF REGION I-REGION II TRANSITION

Earlier, trigger sensitivities were mentioned as being the small voltage and current differences between the turning points of the negative resistance characteristic and the stable operating points. The determination of the turning points and their stability is of great importance since it is usually desired to obtain maximum stable sensitivity. The voltage and current at the two turning points* have been given the subscript p and v for the low and high current conditions respectively as shown in the synopses of Fig. 17, 18 and 19. $V_{\epsilon p}$ and $I_{\epsilon p}$ in the short-

* Sometimes termed "peak" and "valley".

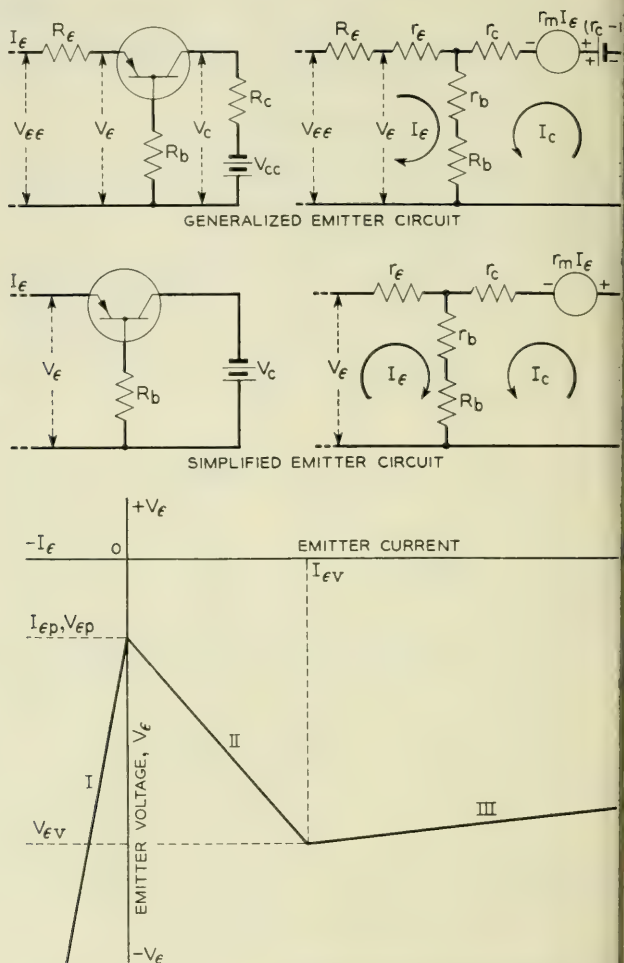


Fig. 17—Synopsis of emitter ne

circuit or $R_e = R_c = 0$ case for example may be obtained by a simultaneous solution of equation (24) for Region I and equation (27) for Region II. Thus

$$V_{\epsilon p} \approx \frac{V_c R_b}{R_b + r_c} \quad (31)$$

and

$$I_{\epsilon p} = 0. \quad (32)$$

That the low current turning point falls on the emitter current axis, i.e., $I_{\epsilon p} = 0$, is a consequence of the original assumption that $r_m = 0$

Synopsis

a)

$$I_{\epsilon} \left[(r_{\epsilon} + r_b + R_b + R_{\epsilon}) - \frac{(r_b + R_b)(r_b + R_b + r_m)}{r_b + R_b + r_c + R_c} \right] \\ \frac{(V_{cc} + I_{cc}(r_c - r_c'''))}{r_b + R_b + r_c + R_c} (r_b + R_b)$$

ximate Short Circuit Case

$$r_b \ll R_b; \quad r_{\epsilon}, r_{\epsilon}''' \ll R_b; \quad R_b \ll r_o, r_m; \quad R_{\epsilon} = R_c = 0; \\ I_{co}(r_c - r_c''') \ll V_c$$

a I

$$V_{\epsilon} \approx I_{\epsilon} r_{\epsilon}' + \frac{V_c R_b}{R_b + r_c}$$

a II

$$V_{\epsilon} \approx I_{\epsilon} R_o(r - \alpha) + \frac{V_c R_b}{R_b + r_c}$$

a III

$$V_{\epsilon} \approx I_{\epsilon}(r_{\epsilon}''' + r_c''' - r_m'') + V_c$$

$$I_{\epsilon p} \approx 0; \quad V_{\epsilon p} = \frac{V_c R_b}{R_b + r_c}$$

$$I_{\epsilon v} = \frac{V_c}{R_b(1 - \alpha)}; \quad V_{\epsilon v} = V_c \left[1 + \frac{r_{\epsilon}''' + r_c''' - r_m''}{R_b(1 - \alpha)} \right]$$

$$\frac{V_{\epsilon v}}{V_{\epsilon p}} = \frac{R_b + r_c}{R_b}$$

once characteristic and properties.

for $I_{\epsilon} < 0$ and $r_m > 0$ for $I_{\epsilon} > 0$. This is not a precise assumption and the turning point will usually lie slightly in the positive emitter current region. For very small triggers or more accurate calculations, consideration must be given to closer approximations of $r_m = f_1(I_{\epsilon})$ and $R_{11} = f_2(I_{\epsilon})$.

The consequences of equation (31) can be quite serious. V_c and R_b are of course fixed, but r_c is variable from unit to unit, with temperature and perhaps with life. The variability of $V_{\epsilon p}$ can result in failure to trigger, self-triggering or lock-up at high current.

EVALUATION OF REGION II-REGION III TRANSITION

The high-current turning point for the short-circuit case is determined from a simultaneous solution of the pertinent equations for Regions II and III, equations (27) and (30). Thus,

$$I_{\epsilon v} \approx \frac{V_c r_c}{R_b(1 - \alpha)(r_c + R_b)} \quad (33)$$

$$V_{\epsilon v} \approx V_c \left[1 + \frac{r_c(r_{\epsilon}''' + r_c''' - r_m''')}{(r_c + R_b)R_b(1 - \alpha)} \right] \quad (34)$$

Where it may be approximated that $r_c \gg R_b$, as has already been done in bringing in α , equations (33) and (34) become,

$$I_{\epsilon v} \approx \frac{V_c}{R_b(1 - \alpha)} \quad (35)$$

$$V_{\epsilon v} \approx V_c \left[1 + \frac{r_{\epsilon}''' + r_c''' - r_m'''}{R_b(1 - \alpha)} \right] \quad (36)$$

In this order of approximation, $V_{\epsilon v}$ is nearly equal to V_c . Any variation in the lower trigger point is primarily with $I_{\epsilon v}$, due chiefly to any change in α . It is interesting to note that the trigger point will move along the Region III curve given by (30).

The ratio of $V_{\epsilon v}$ to $V_{\epsilon p}$ is often of interest to estimate voltage swings or perhaps as a design objective in some switching circuits. Thus from (36) and (31),

$$\frac{V_{\epsilon v}}{V_{\epsilon p}} = \frac{[R_b(1 - \alpha) + r_{\epsilon}''' + r_c''' - r_m'''](R_b + r_c)}{R_b^2(1 - \alpha)} \quad (37)$$

If $r_{\epsilon}''' + r_c''' - r_m''' \ll R_b(1 - \alpha)$ then (37) becomes:

$$\frac{V_{\epsilon v}}{V_{\epsilon p}} = \frac{R_b + r_c}{R_b} \quad (38)$$

If R_b is very large, the ratio approaches unity with the implication of the existence of only two regions. This is equivalent to saying that the negative resistance becomes infinite over an infinitely short range. The proper choice of R_b in terms of (38) may well be a design problem where it is desired to have a high ratio of $V_{\epsilon p}$ to $V_{\epsilon v}$, as in lockout circuits.

SYNOPSIS OF NEGATIVE RESISTANCE CHARACTERISTICS

Synopsis for the three negative resistance characteristics are given in Figs. 17, 18 and 19. The solution and analysis procedures are the same as outlined for the emitter characteristic. It should be noted that the base characteristic is short-circuit stable in distinction to the emitter

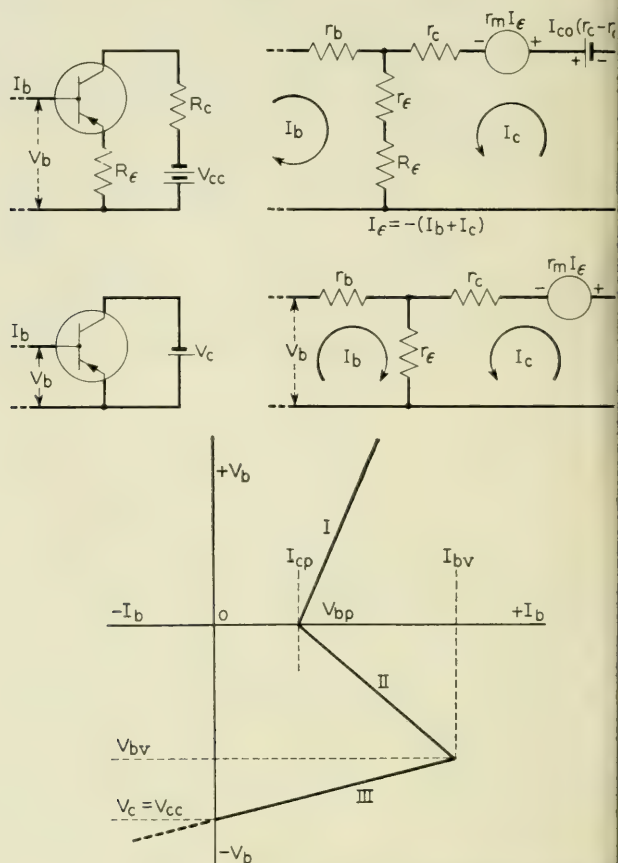


Fig. 19—Synopsis of base ne

and collector characteristics which are open-circuit stable. It would have been more appropriate to solve the base circuit in terms of conductances rather than resistances. The magnitudes of negative resistance obtained in this connection are quite low which may be misleading; the negative conductance is quite high, however, which is desired in short-circuit stable negative resistance circuits.

Care should be taken in the literal employment of the approximate regional relationships in Figs. 17, 18 and 19. They are very definitely approximate and are intended to illustrate behavior and the limiting condition only to bring out the relative importance of device parameters. It is suggested that calculations be started with the general case and approximations be made as are valid. For example, the con-

Synopsis

$$V_b = I_b(r_b + R_b + r_e + R_e) + I_c(r_e + R_e)$$

$$V_{ce} + I_{co}(r_c - r_c'') = I_b(r_e + R_e - r_m) + I_c(r_e + R_e + R_c - r_m)$$

$$I_b \left[r_b + R_b + R_e + r_e - \frac{(r_e + R_e)(r_e + R_e - r_m)}{r_e + R_e + r_c + R_c - r_m} \right] + \frac{\{V_{ce} + I_{co}(r_c - r_c'')\}(r_e + R_e)}{r_e + R_e + r_c + R_c - r_m}$$

Approximate Short Circuit Case

$$R_e = R_c = 0; \quad I_{co}(r_c - r_c'') \ll V_c; \quad r_e \ll r_c(1 - \alpha)$$

Region I

$$V_b \approx I_b \left(\frac{r_e' r_c}{r_c + r_e'} \right) + \frac{V_c r_e'}{r_e' + r_c}$$

Region II

$$V_b \approx I_b \left(\frac{r_b + r_e}{1 - \alpha} \right) + \frac{V_c r_e}{r_c(1 - \alpha)}$$

Region III

$$V_b \approx I_b r_b''' + V_c$$

$$I_{bp} \approx \frac{V_c}{r_c}; \quad V_{bp} = 0$$

$$I_{bv} \approx V_c \left[\frac{1 - \alpha}{r_e} \right]; \quad V_{bv} \approx V_c \left(1 - \frac{(\alpha - 1)r_b'''}{r_e} \right)$$

Characteristic curves and properties.

Conclusion is reached in the collector characteristic that the negative resistance (Region II) is independent of the base resistance or feedback. This is true for only the limited range where $r_e \ll R_b \ll r_c$.

EXAMPLE OF CALCULATED AND EXPERIMENTAL CHARACTERISTICS

An example to illustrate the analysis is shown in Fig. 20 where both experimental and calculated characteristics for the emitter circuit are given. In this example there is appreciable load resistance; hence r_c''' , r_e''' and r_m''' are of no consequence since they will all be very small compared to the R_c of 2.2K ohms. Also, $R_b = 6.8K$ ohms is much greater than r_b ; hence r_b can be neglected. Since V_c is -45 volts, the I_{co} term may also be neglected.

Computing V_{ep} first,

$$V_{ep} = \frac{V_c R_b}{R_b + r_c + R_c} = \frac{-45(6.8K)}{(6.8K + 19K + 2.2K)} = -10.9 \text{ volts} \quad (39)$$

The calculated value of -10.9 volts compares quite favorably to the measured -11.0 volts.

Region II is given in this case, approximately by,

$$V_e \approx \left[R_b + \frac{R_b(R_b + r_m)}{R_b + r_c + R_c} \right] I_e + \frac{V_c R_b}{R_b + r_c + R_c} \quad (40)$$

$$\text{or} \quad V_e \approx \left(6.8K + \frac{6.8K(6.8K + 50K)}{6.8K + 19K + 2.2K} \right) I_e - 10.9 \quad (41)$$

$$V_e \approx (-8.9K)I_e - 10.9 \quad (42)$$

The first term is of course the slope in Region II and is the magnitude of the negative resistance. The calculated value is -8900 ohms whereas the measured value was approximately -9200 ohms.

The Region III approximation, derived also from the general relationship is,

$$V_e = \frac{(R_b R_c)}{R_b + R_c} I_e + \frac{V_c R_b}{R_b + R_c} \quad (43)$$

$$= \frac{((6.8K)(2.2K))}{(6.8K + 2.2K)} I_e = \frac{45(6.8K)}{6.8K + 2.2K} \quad (44)$$

$$\text{or} \quad V_{eIII} = (1785)I_e - 34 \quad (45)$$

The relation for Region III agrees quite well in slope but not in dc value as may be seen in Fig. 20. Since in this example the Region III behavior is determined essentially by the circuit parameters, it is surmised that the nominal 45-volt battery employed in taking the data was actually 47 volts.

The Region I check is essentially perfect since the approximation given in Fig. 17 is quite good.

Note the error at the intersection of the Regions II and III. The broken-line method predicts a sharp transition whereas the actual case is gradual. The deviation is due to the gradual changes in r_m and r_c as the collector voltage approaches cutoff and is the largest gross error in the approximation.

It is believed that analysis of this sort will reasonably predict circuit behavior and lead to device requirements. There must be a thorough understanding of the approximations involved and the accuracy will be directly related to the degree to which the original idealized characteristics are approximated. Extended, by means of more than three broken

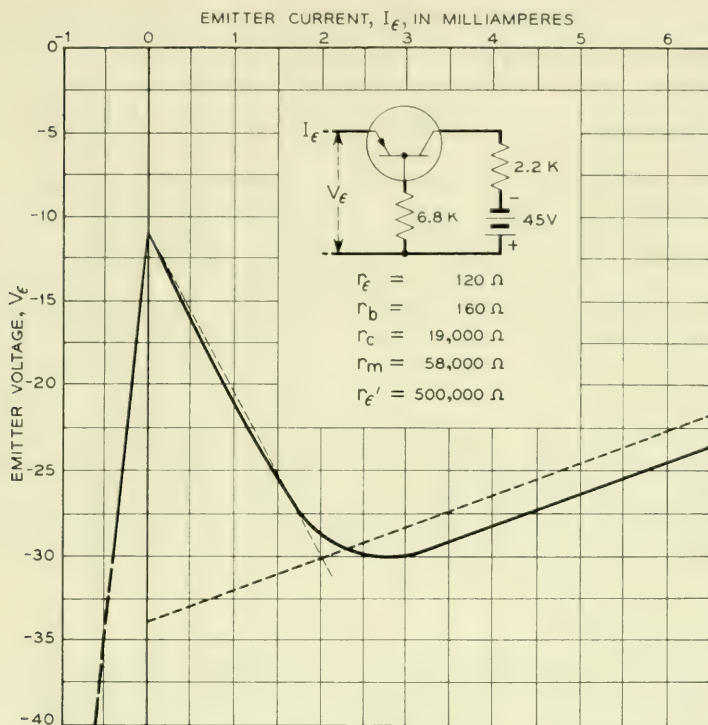


Fig. 20—Experimental and calculated emitter negative resistance characteristic.

lines, the method will yield fine detail to the degree to which device parameters are known and patience will permit. Transient behavior and analysis have not been discussed and are needed for a more complete understanding, particularly where transitional speeds are of concern.*

III—SWITCHING TYPE TRANSISTOR PROPERTIES

An examination of the circuit approximations given in Figs. 17, 18, and 19 will reveal that the transistor and circuit designers will want to know nearly all there is to know about the device characteristics. This is not particularly surprising since the device is used over its entire range rather than over a limited portion as in the case of small-signal applications. The same examination of the circuit relations will also show that

* A treatment of the transient behavior between regions is given in B. G. Farley, "Dynamics of Transistor Negative Resistance", *Proc. Inst. Radio Engr.*, Nov., 1952. Analysis and the solution for the periods of the monostable and astable cases, assuming infinite region to region transition speed, are given in G. E. McDuffie, Jr., "Pulse Duration and Repetition Rate of a Transistor Multivibrator", *Proc. Inst. Radio Engr.*, Nov., 1952.

virtually all of the device parameters should be constant from unit to unit and with ambient conditions.

It can be shown that for small-signals a device may be uniquely characterized by five measurements. In terms of the parameters used here these might be R_{11} , R_{12} , R_{22} , R_{21} and the dc bias point or equally, r_e , r_b , r_c , r_m and the bias point. Since the problem was linearized in the approximation, it follows that 15 such measurements, five in each region, are necessary for proper switching device characterization. The indicated extensive testing required may be reduced somewhat by suitable approximations. It is clear that the switching device designer and producer must reconcile themselves to making more tests for accurate characterization than when small-signal devices are concerned.

What will be given here is a description of typical developmental switching transistors in terms of the parameters which have evolved as a result of practical approximations. The method will be to discuss device properties and measurements region by region; then to discuss the properties at the transition points. Temperature, frequency and life behavior will be taken up separately.

REGION I PROPERTIES

In Region I, the emitter current is negative. Hence the emitter resistance r'_e is large and is essentially that of a diode in the reverse direction. At present r'_e is measured by a simple dc test of the current which flows at a nominal -10 volts. Both r'_e and r_b will be discussed further under the Region I-Region II transition properties.

The Region I collector resistance is one of the most important parameters in switching. This is because of its determining nature in the turning point voltages in Figs. 17, 18, and 19. Actually, what is of concern is not the small-signal slope shown as r_c^* in Fig. 21, but rather the dc current and voltage relationship shown as r_{c0} . For example in Fig. 17, it may be seen that V_{ep} is given by the voltage drop determined by the product of R_b and the dc collector current.

Fig. 21 is an idealization of the R_{22} characteristic and has been designed to bring out the diode nature of the collector by emphasizing the saturation current and voltage, I_{c0} and V_{c0} . In junction devices the break in the $I_c = 0$ characteristic at I_{c0} is quite evident whereas in present point contact devices the transition is smooth due to the much lower values of r_c . The device significance is the same, however; I_{c0} varies rapidly with temperature whereas r_c varies at a considerably lower rate.

* The actual parameter is of course R_{22} , but since $R_{22} = r_e + r_b$ and $r_b \ll r_e$, R_{22} is taken as r_e .

In junction devices the proper measurements would be of I_{c0} and r_c . Since I_{c0} is difficult to define in point contact devices, r_{c0} has been measured as an approximation. In the idealization, r_c and r_{c0} are related as,

$$r_{c0} \doteq \frac{(I_c - I_{c0})}{I_c} r_c \quad (46)$$

The measurement of r_{c0} is made at a collector voltage which is typical of the applications in the range of perhaps -10 to -45 volts.

A constant dissipation line has been drawn on Fig. 21, which reveals the desirability of having r_{c0} very large in order to operate at higher voltages and to secure high efficiency through lower dissipation in the OFF or rest condition.

REGION II PROPERTIES

The Region II low frequency properties are essentially identical to those of transistors intended for small-signal applications. A possible exception is the somewhat less attention paid to the base resistance, r_b , which is critical to small-signal applications. The characterization consists of a normal small-signal set plus dc bias values.

REGION III PROPERTIES

The Region III properties have been defined largely by a figure of

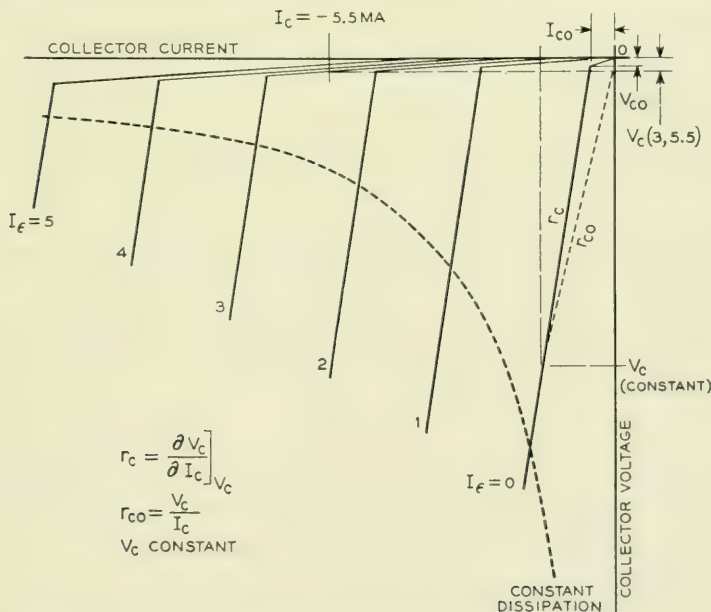


Fig. 21—Idealized output characteristic illustrating parameters.

merit measurement shown as $V_c(3, -5.5)$ in Fig. 21. This measurement is the voltage from collector to base under the condition that $I_e > -(I_c/\alpha)$. In this instance I_e has been chosen to be 3 mA and I_c to be -5.5 mA. The collector current value is chosen on the basis of the smallest tolerable value of alpha expected so as to place the point of measurement near the R_{22} knee, but in Region III or overload.

The $V_c(3, -5.5)$ measurement is a good measurement for defining the general behavior. $V_c(3, -5.5)$ taken with the r_{c0} measurement constitute a very good defining set for checking the transistor as in re-measuring. For design purposes, the $V_c(3, -5.5)$ measurement is not sufficient. It provides an approximate value for r_c''' , but does not define r_e'' and r_m''' . A second dc measurement, the collector to emitter voltage drop, V_{ec} , has been employed experimentally also. An improved characterization will undoubtedly involve separate measurements of r_e''' , r_m''' and r_c''' .

REGION-TO-REGION TRANSITION PROPERTIES

The transition between Regions I and II is accompanied by abrupt changes in r_e and r_m .

The theory assumes that both of these parameters change at an infinite rate at a fixed emitter current, taken as $I_e = 0$. Unfortunately neither of these assumptions is strictly true. r_e undergoes a gradual change from high to low values which is only approximated by the three assigned values. In particular the behavior near $I_e = 0$ is of concern when dealing with small triggers.

The forward transfer impedance changes at a finite rate also. Further, the emitter current at which the maximum rate of change occurs will vary from unit to unit. Present practice also has been to measure α rather than r_m . The rationale for doing so is not too good since r_m is quite likely the better parameter to characterize. Alpha has a strong physical appeal, fits well into the circuit problems and is easy to measure.

Since $\alpha = (r_b + r_m)/(r_b + r_c)$ it is necessary to assume that r_b and r_c are constant near $I_e = 0$, an only fair approximation. Having made the approximation, the typical α behavior shown in Fig. 22 may be taken as a measure of r_m . Three values are measured, the first of which, α_1 , in Region II, is redundant to the Region II small-signal measurements. The two limits, α_2 and α_3 , serve to place lower and upper limits on the absolute values of α at the Regions I-II transition. These limits in turn place a lower value on the rate of change in α within the $I_e \pm \Delta$ range shown.

It may be noted that α in Region I is finite. There is a lower limit

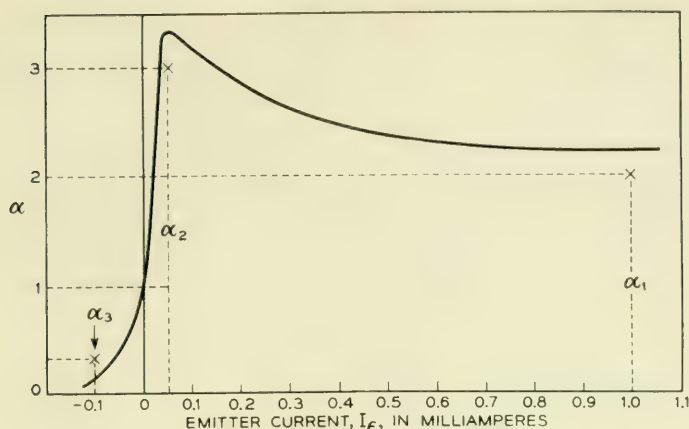


Fig. 22—Alpha characteristic.

even though r_m is zero since $\alpha_1 \rightarrow (r_b/r_b + r_c)$. The values normally encountered at $I_e = 0 - \Delta$ are usually in excess of this lower limit.

The feedback resistance r_b tends to rise as $I_e \rightarrow 0$ which may be important to some trigger circuits. As the circuitry becomes more sophisticated, it is expected that more attention will need to be paid to the behavior of r_e , r_m and r_b at and near $I_e = 0$.

The transition from Region II to Region III is determined from the relation $I_e = -(I_c/\alpha)$. The problem is quite similar to the control of the μ factor in tubes where plate current cut-off is given by $V_g = -(V_p/\mu)$. Present practice has been to depend upon the α_1 values and upon the lower limit placed on alpha in the $V_c(3, -5.5)$ measurement. Further effort is needed here also.

TYPICAL PARAMETER VALUES AND DISTRIBUTIONS

Integrated distribution curves for the parameters of a typical developmental switching transistor are shown in Fig. 23. The unit-to-unit variations are deemed to compare favorably with those of commercial electron tubes. The parameter of most serious variability is r_{c0} which is unfortunate since r_{c0} is so important to trigger sensitivity stability.

TEMPERATURE, FREQUENCY AND SHOCK PROPERTIES

Transistor parameters are reasonably constant with temperatures below room temperature. Above room temperatures some of the parameters are variable. r_e and r_b are fairly constant, changing very little to 70°C . r_c and r_m decrease fairly rapidly, maintaining a ratio such that alpha rises slightly. r'_e and r_{c0} change most rapidly and, while both of

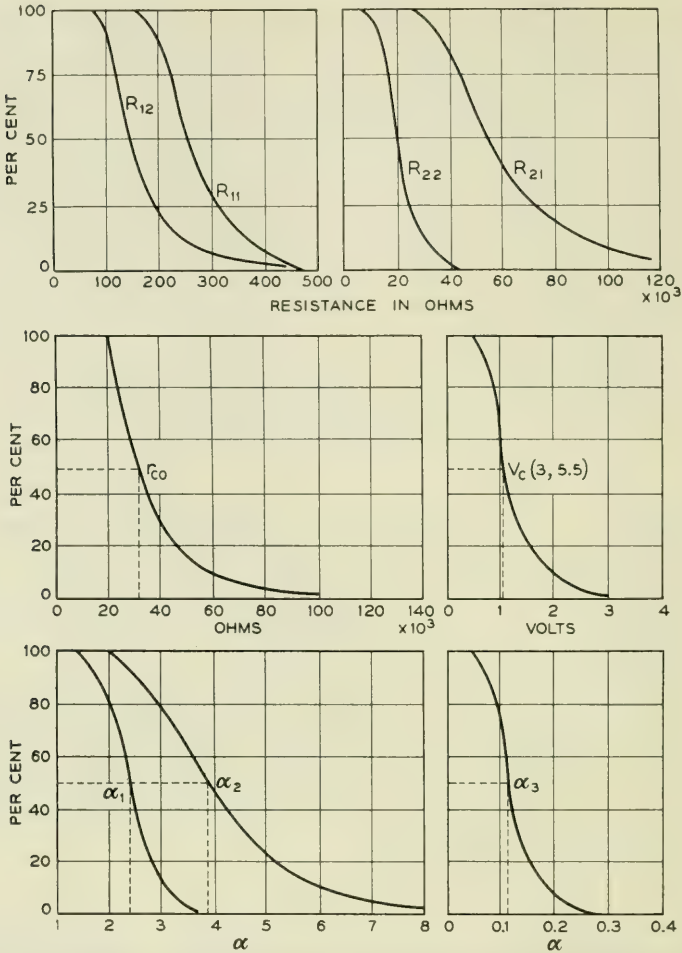


Fig. 23—Variation in parameters of developmental switching type transistor (M1698).

these parameters are of little consequence in small-signal applications, they are quite important in switching, particularly r_{c0} .

Early transistors might exhibit a change in r_{c0} at 60°C of 3 to 1 or more from room temperature values. The transistors of which the data in Figs. 13 and 23 are typical have an r_{c0} temperature coefficient of about -0.75 per cent/°C. That is, the room temperature value of r_{c0} might be reduced by 30 per cent at 70°C. The improved temperature behavior implies a corresponding reduction in variation in trigger sensitivity. Parameter values, large-signal and small-signal, are shown in Fig. 24 as a function of temperature.

Variation in characteristics will arise from self-engendered heat, that is, dissipation. Transistors may be thermally unstable under constant voltage conditions. Since the switching properties are exhibited under short-circuit or constant voltage terminations, thermal properties are of concern. The limitations involved are similar to those of any positive feedback circuit. If the thermal loss through radiation and conduction exceeds the heat input, the system will be stable. The practical significance is to place limitations on dissipation and to employ designs which result in rapid heat loss. Other design criteria such as miniaturization may limit the latter.

If perfect switching characteristics were obtainable, dissipation would be of little consequence in switching. This is akin to saying that neither a short-circuit nor an open-circuit dissipates any energy. Further, the perfect device has zero transition time and therefore involves no loss. The transistor has finite resistance both open and closed and a finite although rapid transition time. There is some advantage however. A constant dissipation curve shown as a dotted line has been included in Fig. 21. Small-signal operation at mid-range currents and voltages results in fairly low limitations on both current and voltage. The intersection with the R_{22} voltage saturation line ($I_e = 0$) is at fairly high voltage. Similarly, the intersection with the collector voltage cut off line is at high current. For constant dissipation, approximately,

$$\text{Voltage saturation:} \quad P_d \approx \frac{V_c^2}{r_{c0}}$$

$$\text{Voltage cutoff:} \quad P_d \approx I_c^2(r_e''' + r_c''' - r_m''')^*$$

Depending upon the circuit the assumed dissipation limit may or may not be exceeded during the transitions. Should the limit be exceeded,

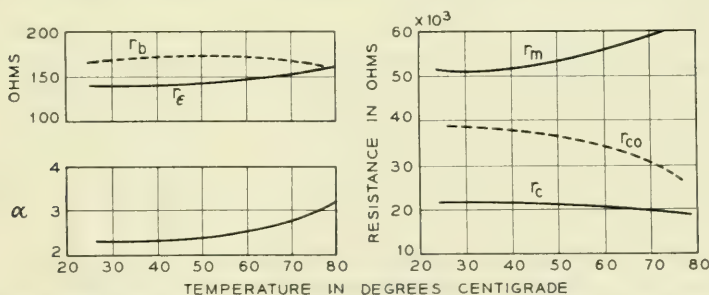


Fig. 24—Temperature behavior of characteristics of developmental switching-type transistor (M1689).

* This includes both emitter and collector dissipation. See equation (30).

and substantially so, there are normally no serious consequences due to the very rapid transitions and consequent low thermal energy generated.

Transistors may not be able to tolerate excess dissipation on this basis if the circuits are slow, that is with transition times in excess of perhaps a few tenths of a microsecond. Such conditions may arise, for example, if loads are inductive. In many such cases, shunting capacitor networks will often permit a rapid transition with consequent transfer of current to the inductive load.

The frequency response of point contact transistors can be sufficiently good to insure switching type operation with rise times of the order of 0.1 to 0.01 μ s. Fall times may be somewhat longer due to the hole storage effect. In regenerative circuits, operating speeds are faster than might be imagined from the small-signal frequency cutoff. Reliable operation with rise times of 0.1 μ s is obtained with only nominal attention to frequency cutoff. Speeds of the order of 0.02 μ s require a 10 mc. lower limit. Present junction transistors are substantially slower.

Accurate life estimates are difficult to make due to the rapid rate of development, the relative age of the transistor and the number of parameters involved. A given device is quite likely to be obsolete and forced to give way to an improved version before sufficient models can be obtained for life tests. A small quantity of transistors having properties similar to those of Fig. 20 and 21 have been operated for over 6,000 hours with an indicated life of 30,000 hours. Other similar transistors with longer life histories have indicated lives of better than 70,000 hours. The pattern appears to be similar to that of electron tubes—an early failure and change rate followed by a very slow exponential rate. It is believed that life is extended by low power operation and is decreased by high temperature operation.

The relatively high noise level of transistors does not appear to be a significant problem at present when considered in terms of automata. Systems employing switching type circuits in pulse communication will of course be concerned. It is suggested that the non-concern for noise in non-transmission type systems is largely a reflection of the ease with which high magnitudes of state changes are obtained. With design trends toward low power and low operating levels, noise will undoubtedly set a lower limit of level operation in such systems also.

The extreme resistance of the transistor to shock and vibration with a consequent absence of microphonism may in some applications result in effective lower noise. Shocks in excess of 20,000G have resulted in no damage. No evidences of current modulation in excess of noise have been detected with vibrational forces of the order of 100G at frequencies

as high as 1000 cycles in tests on the transistor of Fig. 1. Transistors have been included in plastic embedded circuits without change of characteristics.

SUMMARY—TRANSISTOR PROPERTIES

Transistors have been designed with properties expressly intended for switching applications.* The characteristics are acceptable for contemporary switching type circuits and sufficiently reproducible to permit interchangeability of devices in circuits of normal requirements. The characterization has been sufficiently unique to permit the calculation of first order circuit performance. The characterization is not sufficiently complete to permit determination of the complete transient behavior.

In terms of the circuits described, the major parameter limitation is concerned with the variability of the d-c collector resistance among units and with temperature. It is expected that future circuit development will place additional requirements on the transistor, particularly as related to the transitions between regions. It is also to be expected that future circuit designs may establish new or modify present device requirements.

A major consideration for computer or computer-like systems, reliability, particularly with respect to time and temperature, has not been established, but appears to be favorable.

ACKNOWLEDGMENT

It is impossible to properly acknowledge credit to all of those who contributed to the concepts, data and results of this paper. Particular acknowledgment is due to J. A. Morton who provided first the method of attack for the analysis and second, continued stimulation. Acknowledgment is also given to A. J. Rack who first classified and explained the several simple circuits of the first section. J. J. Kleimack provided transistor data and R. L. Trent, circuit data.

* See Reference 3 also.

Abstracts of Bell System Technical Papers* Not Published in This Journal

An Approximate Quantum Theory of the Antiferromagnetic Ground State. P. W. ANDERSON¹. *Phys. Rev.*, **86**, pp. 694–701, June 1, 1952. (Monograph 1995).

A careful treatment of the zero-point energy of the spin-waves in the Kramers-Heller semiclassical theory of ferromagnetics leads to surprisingly exact results for the properties of the ground state, as shown by Klein and Smith. An analogous treatment of the antiferromagnetic ground state, whose properties were unknown, is here carried out and justified. The results are expected to be valid to order $1/S$ or better, where S is the spin quantum number of the separate atoms.

The energy of the ground state is computed and found to lie within limits found elsewhere on rigorous grounds. For the linear chain, there is no long-range order in the ground state; for the simple cubic and plane square lattices, a finite long-range order in the ground state is found. The fact that this order can be observed experimentally, somewhat puzzling since one knows the ground state to be a singlet, is explained.

Method of Synthesis of the Statistical and Impact Theories of Pressure Broadening. P. W. ANDERSON¹. Letter to the Editor. *Phys. Rev.*, **86**, p. 809, June 1, 1952.

Arcing at Electrical Contacts on Closure. Part III. Development of an Arc. L. H. GERMER¹ and J. L. SMITH¹. *Jl. Applied Phys.*, **23**, pp. 553–562, May, 1952. (Monograph 2002).

A description is given of a system made up of experimental electrodes and an oscilloscope by means of which the potential across the electrodes can be recorded with a time resolution of about 10^{-9} sec. and a potential sensitivity of 1-trace width per volt. The closure of the electrodes to produce a short arc is synchronized with the oscilloscope sweep so that the beginning of the arc is photographed.

As an arc starts the potential across the electrodes decreases more or less gradually from the applied voltage to a steady value characteristic of the metal of the electrodes. The course of this change is extremely variable as is also the time over which the change is spread. The average value of the time appears to

* Certain of these papers are available as Bell System Monographs and may be obtained on request to the Publication Department, Bell Telephone Laboratories, Inc., 463 West Street, New York 14, N. Y. For papers available in this form, the monograph number is given in parentheses following the date of publication, and this number should be given in all requests.

¹ Bell Telephone Laboratories

vary with circuit inductance and with the nature of the electrode surfaces. For inactive silver surfaces and an inductance of $0.10 \mu h$ the average value of the time is about $0.007 \mu \text{ sec.}$, and for active surfaces and the same inductance $0.011 \mu \text{ sec.}$ For active surfaces and an inductance of $5 \mu h$ the average value of the time is $0.02 \mu \text{ sec.}$

The electrode separation at which an arc strikes is determined from the oscilloscope traces and from a correction for the height of the mound of metal thrown up by the arc. For active silver electrodes the average separation (at 40 or 45 volts) corresponds to a gross electric field of $0.8 \times 10^6 \text{ volts/cm.}$ and for inactive silver electrodes to a field of $2.3 \times 10^6 \text{ volts/cm.}$ These are probably better values than earlier measurements of these fields. There has not yet been any success in interpreting these phenomena in terms of fundamental processes.

A Carrier Telegraph System for Short-Haul Applications. J. L. HYSKO¹, W. T. REA¹ and L. C. ROBERTS¹. *Elec. Eng.*, **71**, pp. 625-630, July, 1952. (Monograph 2006).

This compact frequency-shift carrier telegraph system provides channels in and above the voice range. The channel terminal unit incorporates arrangements for handling Teletypewriter Exchange Service supervisory signals and employs no electromagnetic relays.

Some Problems in Sampling Accounting Procedure. H. L. JONES¹. pp. 209-250. *Am. Soc. for Quality Control*. Quality control conference papers. 6th Annual Convention. N. Y., Am. Soc. Quality Control, 1952.

Photometric Determination of Beryllium in Beryllium-Copper Alloys. C. L. LUKE¹ and M. E. CAMPBELL¹. *Anal. Chem.*, **24**, pp. 1056-1057, June, 1952. (Monograph 2013).

Steady Rotational Flow of Ideal Gases. R. C. PRIM¹. *Jl. Rational Mech. and Analysis*, **1**, pp. 425-497, July, 1952.

This paper concerns the steady rotational flow of non-viscous and thermally non-conducting gases subject to no extraneous force field. For the most part attention is restricted to gases having constant specific heats. However, some of the results are valid for more general classes of fluids. Uniformity of total flow energy (stagnation enthalpy) or of entropy throughout the flow is not assumed. The present work is intended to be a comprehensive treatment of the status (in 1949) or rotational flow theory from the point of view of the establishment of general properties of such flows and the discovery and study of families of exact solutions to the equations governing them.

Finishing Metal Parts for Telephones. F. B. RINCK³. *Metal Progress*, **61**, pp. 65-70, June, 1952.

¹ Bell Telephone Laboratories

³ Western Electric Company

Binary Counter Uses Two Transistors. R. L. TRENT¹. *Electronics*, **25**, pp. 100-101, July, 1952.

Various timing and registry functions are provided by transistorized counter with repetition rate from 0 to 50 kc. It has stability without the usual sacrifice in sensitivity and it permits either positive or negative triggering pulses to be used.

Structural Imperfections in Quartz Crystals. W. L. BOND¹ and J. ANDRUS¹. *Am. Mineral.*, **37**, pp. 622-632, July-August, 1952. (Monograph 2001).

A method for examining the topography of atomic planes is developed and applied to quartz crystals. It is thought to have higher resolution than the method of Wooster and Wooster (*Nature*, **155**, p. 786 (1945)), or that of Ramachandran (*Proc. Ind. Acad. Sci.*, **19A**, p. 280 (1944)). Because of the higher resolution it gives more detailed information. A fair percentage of ostensibly perfect quartz is shown to have slight irregularities.

Packaging Principles Employing Plastics and Printed Wiring to Improve Reliability. W. J. CLARKE¹ and N. J. EICH¹. pp. 133-137. A. I. E. E., I. R. E. and R. T. M. A. Symposium, Progress in Quality Electronic Components. Proceedings, Wash., D. C., May 5-7, 1952. Wash., D. C., R. T. M. A., 1952.

Miniaturized Components for Transistor Action. P. S. DARNELL¹. pp. 51-57. A. I. E. E., I. R. E. and R. T. M. A. Symposium, Progress in Quality Electronic Components. Proceedings, Wash. D. C., May 5-7, 1952. Wash., D. C., R. T. M. A., 1952.

Some Basic Concepts of Quality Control. G. D. EDWARDS¹. Shewhart Medalist Address. *Ind. Quality Control*, **9**, pp. 9-10, July, 1952.

Effective Sum of Multiple Echoes in Television. A. D. FOWLER¹ and H. N. CHRISTOPHER¹. *S. M. P. T. E., JI.*, **58**, pp. 491-500, June, 1952.

Observers compared the interfering effect of multiple echoes with that of single echoes in black-and-white television pictures. The multiple echoes were 2, 4 or 8 echoes of equal strength but different delays. The single echoes were 40, 35 or 30 db weaker than the main signal. A method for estimating addition effects of several echoes is presented and demonstrated to be consistent with the test results.

Design Factors Influencing the Reliability of Relays. J. R. FRY¹. pp. 101-107. A. I. E. E., I. R. E. and R. T. M. A. Symposium, Progress in

¹ Bell Telephone Laboratories

Quality Electronic Components. Proceedings, Wash., D. C., May 5-7, 1952. Wash., D. C., R. T. M. A., 1952.

Energy of a Bloch Wall on the Band Picture. II. Perturbation Approach. C. HERRING¹. *Phys. Rev.*, **87**, pp. 60-70, July 1, 1952.

The "exchange stiffness" constant, which appears in the theory of the Bloch interdomain wall in ferromagnetics, can be calculated by computing the response of a saturated specimen to a small spatially varying perturbing field. This calculation is carried out here in the self-consistent field approximation, using running waves for the one-electron states, and the result is interpreted physically in terms of precession of the spins of moving electrons. Combination of the present theory with the Stoner-Wohlfarth model of the ferromagnetic electrons in nickel does not give satisfactory results, probably because the latter model does not approximate the actual self-consistent field solution very well. However, application of the theory to the free electron gas is of interest as a confirmation of the validity of the perturbation approach. It is shown that there exist, even in a ferromagnetic metal, quantum states orthogonal to all the low-lying states of the conventional band picture and having the properties of spin waves. The presumably universal relation between the exchange stiffness constant and the energies of spin waves of long wavelength is verified in the present approximation. It is shown that spin waves carry a current in a metal, though not in an insulator. For spin waves of long wavelength the present theory can be shown to include Slater's theory of spin waves in a ferromagnetic insulator, and a fortiori to include all previous theories based on the atomic model.

Nonsynchronous Pulse Multiplex System. A. L. HOPPER¹. *Electronics*, **25**, pp. 116-120, August, 1952.

Voice transmitters use one frequency simultaneously but no synchronizing pulse is necessary, although time-division multiplexing is used. Random samples from each transmitter are tagged for identification at proper receiver. System is applicable to rural telephony and moving-vehicle communication.

Design of Modulation Equipment for Modern Single-Sideband Transmitters. A. E. KERWIEN¹. *I. R. E., Proc.*, **40**, pp. 797-803, July, 1952. (Monograph 2012).

This paper deals with considerations that go into the design of modulation equipment for a single-sideband radiotelephone transmitter in which filters are used for sideband suppression. Balance requirements, frequency stability, the choice of intermediate frequencies, and methods of avoiding transmission of spurious frequencies are among the factors which are discussed.

A Multichannel Single-Sideband Radio Transmitter. L. M. KLENK¹, A. J. MUNN¹, and J. NEDELKA¹. *I. R. E. Proc.*, **40**, pp. 797-803, July, 1952. (Monograph 2012).

This paper describes a new single-sideband radio transmitter for transoceanic

¹ Bell Telephone Laboratories

service which represents a substantial improvement over past design. Its important features include: (a) a frequency band which permits deriving four telephone channels, if desired; (b) a push-button method for changing frequencies within a matter of seconds; (c) an increase in power over its predecessor; and (d) all-around improved transmission performance.

Photometric Determination of Aluminum in Lead, Antimony, and Tin and Their Alloys. C. L. LUKE¹. *Anal. Chem.*, **24**, pp. 1122–1126, July, 1952. (Monograph 2013).

The work was undertaken because of a need for a reliable method for the determination of traces of aluminum in lead, antimony, and tin and their alloys. As a preparatory step toward the development of such a method, a thorough study of the specificity of the aluminon-thioglycolic acid and the oxine-cyanide-peroxide photometric aluminum methods was made. As a result, an accurate specific method for aluminum has been developed. This method is applicable to the analysis of lead, antimony, and tin and their alloys and can also be adapted for use in the analysis of a wide variety of other ferrous and nonferrous alloys.

Photometric Determination in Manganese Bronze, Zinc Die Casting Alloys, and Magnesium Alloys. C. L. LUKE¹. and K. C. BRAUN⁴. *Anal. Chem.* **24**, pp. 1120–1122, July, 1952. (Monograph 2013).

The work was undertaken in an effort to produce a rapid reliable method for the determination of aluminum appearing as a major constituent in copper, zinc, and magnesium alloys. The work shows that the photometric aluminum method described by Craft and Makepeace is very satisfactory and that by employing thioglycolic acid as a complexing agent it is possible to simplify the usual photometric methods for the determination of aluminum in nonferrous alloys. The paper contains experimental material that will aid future workers in the application of this method to other materials.

Amplifiers for Multichannel Single-Sideband Radio Transmitters. N. LUND¹, C. F. P. ROSE¹ and L. G. YOUNG¹. *I. R. E., Proc.*, **40**, pp. 790–796, July, 1952. (Monograph 2012).

Considerations are given for designing high-frequency amplifiers whose performance will meet the high standards required for amplifying multichannel signals. A relationship between tone and speech data is presented to show how the tone rating of the amplifier can be determined from the speech rating and interchannel modulation noise requirements.

Measurement of Dynamic Shear Viscosity and Stiffness of Viscous Liquids by Means of Traveling Torsional Waves. H. J. McSKIMIN¹. *Acoustical Soc. Am. Jl.*, **24**, pp. 355–365, July, 1952.

A short periodically repeated train of torsional waves is transmitted along a glass or metal cylindrical rod. After reflection from the free end, these waves are

¹ Bell Telephone Laboratories

⁴ American Smelting and Refining Company, Barber, N. J.

sent back to the quartz crystal which serves as both transmitter and receiver. The phase shift and added attenuation caused by immersing the rod in the test liquid are measured by means of a special balancing arrangement, and yield a calculation of the impedance presented to the rod surface. From an analysis of wave propagation both in the rod and in the liquid, one can calculate the characteristic shear impedance of the liquid, and the dynamic viscosity and stiffness. Data for polyisobutylene liquids with static viscosities up to 2000 poises are given for the frequency range 25–150 kc. High frequency data (5–25 mc) for the same liquids obtained by a method previously reported on (see reference 10 (b)) are correlated to the present work. Some results for polypropylene, polyisoprene, polybutadiene, and polypropylene sebacate are also given.

New Transistors Give Improved Performance. J. A. MORTON¹. *Electronics*, **25**, pp. 100–103, August, 1952.

Better manufacturing processes and germanium materials have provided greater reliability and reproducibility and improved frequency response. Higher power output and better noise figure for high-sensitivity applications are properties of new types.

Microwaves. J. R. PIERCE¹. *Sci. Am.*, **187**, pp. 43–51, August, 1952.

They are radio waves that range in length from about a quarter of an inch to two feet. Investigated during the war for their utility in radar, they are now widely applied in communication.

Glass Unit for Liquid and Vapor Phase Extraction Employing a Single Processing Chamber. H. A. SAUER¹. *Anal. Chem.*, **24**, p. 1232, July, 1952.

The Transistors Development Status at Bell Telephone Laboratories, with Demonstration. W. R. SITTNER¹. pp. 138–142. A. I. E. E., I. R. E. and R. T. M. A. Symposium, Progress in Quality Electron Components. *Proceedings*, Wash., D. C., May 5–7, 1952. Wash., D. C., R. T. M. A., 1952.

Polyethylene Terephthalate as a Capacitor Dielectric. M. C. WOOLEY¹, G. T. KOHMAN¹ and W. McMAHON¹. *Elec. Eng.*, **71**, pp. 715–717, Aug., 1952.

Polyethylene terephthalate, or "Mylar", is a new rival of paper for use as the dielectric in electric capacitors. It appears superior in regard to insulation resistance, temperature coefficient of capacitance, and operating temperature range.

¹ Bell Telephone Laboratories

Contributors to this Issue

A. EUGENE ANDERSON, B.Sc. in E.E., Ohio State University, 1939; M.Sc., Ohio State University, 1939. U. S. Army, 1942-46. Bell Telephone Laboratories, 1939-. Mr. Anderson is concerned with the development of semi-conductor devices, including the transistor. In the past he has been engaged in microwave electron tube and electron beam tube development. Member of I. R. E., Tau Beta Pi, Sigma Xi, Eta Kappa Nu, and Sigma Pi Sigma.

ARTHUR C. KELLER. B.S., Cooper Union, 1923; M.S., Yale University, 1925; E.E., Cooper Union, 1926; Columbia University, 1926-30; Western Electric Company, 1917-25; Bell Telephone Laboratories, 1925-. Special Apparatus Development Engineer, 1943; Switching Apparatus Development Engineer, 1946; Assistant Director of Switching Apparatus Development, 1949; Director of Switching Apparatus Development, 1949-. Mr. Keller's experience in the Bell System includes development and design of telephone instruments; development of systems and apparatus for recording and reproducing sound; and, during World War II, the development, design, and preparation for manufacture of sonar systems and apparatus. His department, in addition to being responsible for a number of military projects, is responsible for the fundamental studies of switching apparatus and the development, design, and preparation for manufacture of electromagnetic and electromechanical switching apparatus for telephone systems. Member of the American Physical Society, A. I. E. E., Acoustical Society of America, I. R. E., S. M. P. T. E., and the Yale Engineering Association. Representative for Bell Telephone Laboratories in the Society for Experimental Stress Analysis. For his contributions to the Navy during World War II, he received awards from the Bureau of Ships and the Bureau of Ordnance.

SAMUEL P. MORGAN, JR., B.S., California Institute of Technology, 1943; M.S., California Institute of Technology, 1944; Ph.D., California Institute of Technology, 1947. Bell Telephone Laboratories, 1947-. A research mathematician, Dr. Morgan specializes in electromagnetic theory. He has been particularly concerned with problems of wave guide and coaxial cable transmission. Member of the American Physical Society, Tau Beta Pi, and an associate member of Sigma Xi.

OSCAR MYERS, B. Chem., Cornell University, 1921. Western Electric Company, 1921-24. Bell Telephone Laboratories, 1924-. Mr. Myers' early work was concerned with circuit testing. He then worked in a circuit design group, from 1929 until 1948. Since 1948 he has been a member of the Switching Engineering Department, and has contributed to the design or development of practically all of the switching developments of the Laboratories, particularly in the field of common controls. His work has covered panel, crossbar, automatic message accounting, toll, crossbar tandem, and other systems. Member of A. I. E. E.

W. RAE YOUNG, JR., B.S., in E.E., University of Michigan, 1937; Bell Telephone Laboratories, 1937-. Mr. Young is in the Systems Studies Department, where he is giving consideration to new system possibilities for meeting future communication needs. During World War II, he worked in radar development and, later, on systems problems in radio communications. From 1945-50 Mr. Young helped set up Bell System performance requirements for mobile radio telephone equipment. Member of I. R. E. and Sigma Xi.

Index to Volume XXXI

A

Admittance

Impedance Bridges for the Megacycle Range, *H. T. Wilhelm*, pages 999-1012.

Alphabets

A Comparison of Signalling Alphabets, *E. N. Gilbert*, pages 504-522.

Anderson, A. E., Transistors in Switching Circuits, pages 1207-1249.

Application of Boolean Algebra to Switching Circuit Design, *R. E. Staehler*, pages 280-305.

Armatures, Relay

Relay Armature Rebound Analysis, *E. E. Sumner*, pages 172-200.

Automatic Switching for Nationwide Telephone Service, *A. B. Clark and H. S. Osborne*, pages 823-831.

Automatic Toll Switching Systems, *F. F. Shipley*, pages 860-882.

B

Baker, W. O. and Heiss, J. H., Interaction of Polymers and Mechanical Waves, pages 306-356.

Barium Titanate

New Techniques for Measuring Forces and Wear in Telephone Switching Apparatus, *W. P. Mason and S. D. White*, pages 469-503.

Bridges, Impedance

Impedance Bridges for the Megacycle Range, *H. T. Wilhelm*, pages 999-1012.

C

Cables, Sheaths and Sheath Losses

Principal Strains in Cable Sheaths and Other Buckled Surfaces, *I. L. Hopkins*, pages 523-529.

Carrier Telegraph System for Short-Haul Applications, *J. L. Hysko, W. T. Rea and L. C. Roberts*, pages 666-687.

Circuits, Switching

An Application of Boolean Algebra to Switching Circuit Design, *R. E. Staehler*, pages 280-305.

Circuits, Traveling-Wave-Tube

Propagation Studies at Microwave Frequencies by Means of Very Short Pulses, *O. E. DeLange*, pages 91-103.

Circuits, Trigger-Type

Transistors in Switching Circuits, *A. E. Anderson*, pages 1207-1249.

Clark, A. B. and Osborne, H. S., Automatic Switching for Nationwide Telephone Service, pages 823-831.

Clos, Charles and Wilkinson, R. I., Dialing Habits of Telephone Customers, pages 32-67.

Coding, Communication

Efficient Coding, *B. M. Oliver*, pages 724-750.

Coils, Relay

Important Design Factors Influencing Reliability of Relays, *J. R. Fry*, pages 976-998.

Common Control Telephone Switching Systems, *Oscar Myers*, pages 1086-1120.

Communication Theory

Efficient Coding, *B. M. Oliver*, pages 724-750.

Comparison of Signalling Alphabets, *E. N. Gilbert*, pages 504-522.

Computers

Selective Fading of Microwaves, *A. B. Crawford and W. C. Jakes, Jr.*, pages 68-90.

Contacts

A New General Purpose Relay for Telephone Switching Systems, *A. C. Keller*, pages 1023-1067.

Important Design Factors Influencing Reliability of Relays, *J. R. Fry*, pages 976-998.

Crawford, A. B. and Jakes, W. C., Jr., Selective Fading of Microwaves, pages 68-90.
Crystals, Germanium

Electrical Noise in Semiconductors, *H. C. Montgomery*, pages 950-975.

D

Darlington, Sidney, Network Synthesis Using Tchebycheff Polynomial Series, pages 613-665.

DeLange, O. E., Propagation Studies at Microwave Frequencies by Means of Very Short Pulses, pages 91-103.

Design Factors Influencing Reliability of Relays, Important, *J. R. Fry*, pages 976-998.

Dialing Habits of Telephone Customers, *Charles Clos and R. I. Wilkinson*, pages 32-67.

Dialing, Nationwide

Automatic Switching for Nationwide Telephone Service, *A. B. Clark and H. S. Osborne*, pages 823-831.

Nationwide Numbering Plan, *W. H. Nunn*, pages 851-859.

Dielectric Mismatch

Mathematical Theory of Laminated Transmission Lines, *S. P. Morgan, Jr.*, pages 883-949; 1121-1206.

Direct Dial Control Systems in Switching

Common Control Telephone Switching Systems, *Oscar Myers*, pages 1086-1120.

Dynamics

Relay Armature Rebound Analysis, *E. E. Sumner*, pages 172-200.

E

Edwards, P. G. and Montfort, L. R., The Type-O Carrier System, pages 688-723.

Efficient Coding, *B. M. Oliver*, pages 724-750.

Elasticity

Interaction of Polymers and Mechanical Waves. *W. O. Baker and J. H. Heiss*, pages 306-356.

Elasticity

Mechanical Properties of Polymers at Ultrasonic Frequencies, *W. P. Mason and H. J. McSkimin*, pages 122-171.

Electric Circuit Theory

Network Representation of Transcendental Impedance Functions, *M. K. Zinn*, pages 378-404.

Electric Networks

Introduction to Formal Realizability Theory, *Brockway McMillan*, pages 217-279; 541-600.

Network Synthesis Using Tchebycheff Polynomial Series, *Sidney Darlington*, pages 613-665.

Electrical Noise in Semiconductors, *H. C. Montgomery*, pages 950-975.

Electromagnetism

Mathematical Theory of Laminated Transmission Lines, *S. P. Morgan, Jr.*, pages 883-949; 1121-1206.

Electronically Controlled Automatic Switching System, An Experimental, *W. A. Malthaner and H. E. Vaughan*, pages 443-468.

Experiments with Linear Prediction in Television, *C. W. Harrison*, pages 764-783.

F

Faraday Effect

The Ferromagnetic Faraday Effect at Microwave Frequencies and Its Applications—The Microwave Gyrator, *C. L. Hogan*, pages 1-31.

Ferromagnetic Faraday Effect at Microwave Frequencies and Its Applications—The Microwave Gyrator, *C. L. Hogan*, pages 1-31.

Formal Realizability Theory, Introduction to, *Brockway McMillan*, pages 217-279; 541-600.

Fry, J. R., Important Design Factors Influencing Reliability of Relays, pages 976-998.

Fundamental Plans for Toll Telephone Plant, *J. J. Pilliod*, pages 832-850.

G

General Purpose Relay for Telephone Switching Systems, A New, *A. C. Keller*, pages 1023-1067.

Gilbert, E. N., A Comparison of Signalling Alphabets, pages 504-522.

Graphs

A Comparison of Signalling Alphabets, *E. N. Gilbert*, pages 504-522.

Gyrator

The Ferromagnetic Faraday Effect at Microwave Frequencies and Its Applications—
The Microwave Gyrator, *C. L. Hogan*, pages 1-31.

H

Harrison, C. W., Experiments with Linear Prediction in Television, pages 764-783.

Hayward, W. S., Jr., The Reliability of Telephone Traffic Load Measurements by Switch
Counts, pages 357-377.

Heiss, J. H. and Baker, W. O., Interaction of Polymers and Mechanical Waves, pages
306-356.

Hogan, C. L., The Ferromagnetic Faraday Effect at Microwave Frequencies and Its
Applications—The Microwave Gyrator, pages 1-31.

Hopkins, I. L., Principal Strains in Cable Sheaths and Other Buckled Surfaces, pages
523-529.

Hysko, J. L., Rea, W. T. and Roberts, L. C., A Carrier Telegraph System for Short-Haul
Applications, pages 666-687.

I

Impedance Bridges for the Megacycle Range, *H. T. Wilhelm*, pages 999-1012.

Impedance Functions

Network Representation of Transcendental Impedance Functions, *M. K. Zinn*, pages
378-404.

Impedance Measurements

Mechanical Properties of Polymers at Ultrasonic Frequencies, *W. P. Mason and H. J.
McSkimin*, pages 122-171.

Interaction of Polymers and Mechanical Waves, *W. O. Baker and J. H. Heiss*, pages
306-356.

Ionization

Properties of Ionic Bombarded Silicon, *R. S. Ohl*, pages 104-121.

Iron Oxide

A New Recording Medium for Transcribed Message Services, *J. Z. Menard*, pages
530-540.

J

Jakes, W. C., Jr. and Crawford, A. B., Selective Fading of Microwaves, pages 68-90.

K

Keller, A. C., A New General Purpose Relay for Telephone Switching Systems, pages
1023-1067.

Kingsbury, E. F. and Ohl, R. S., Photoelectric Properties of Ionically Bombarded Silicon,
pages 802-815.

Kretzmer, E. R., Statistics of Television Signals, pages 751-763.

L

Laminated Transmission Lines, Mathematical Theory of, *S. P. Morgan, Jr.*, pages 883-
949.

Lee de Forest and William Shockley Discuss Electronics, page 612.

Level Distribution Recorder

Comparison of Mobile Radio Transmission at 150, 450, 900, 3700 Mc, *W. R. Young,
Jr.*, pages 1068-1085.

Linear Prediction

Experiments with Linear Prediction in Television, *C. W. Harrison*, pages 764-783.

M

Magnetic Materials

Important Design Factors Influencing Reliability of Relays, *J. R. Fry*, pages 976-998.

Magnetic Recording

A New Recording Medium for Transcribed Message Services, *J. Z. Menard*, pages
530-540.

- Malthaner, W. A. and Vaughan, H. E.*, An Experimental Electronically Controlled Automatic Switching System, pages 443-468.
- Manufacturing Processes
- Important Design Factors Influencing Reliability of Relays, *J. R. Fry*, pages 976-998.
- Mason, W. P. and McSkimin, H. J.*, Mechanical Properties of Polymers at Ultrasonic Frequencies, pages 122-171.
- Mason, W. P. and White, S. D.*, New Techniques for Measuring Forces and Wear in Telephone Switching Apparatus, pages 469-503.
- McMillan, Brockway*, Introduction to Formal Realizability Theory, pages 217-279; 541-600.
- McSkimin, H. J. and Mason, W. P.*, Mechanical Properties of Polymers at Ultrasonic Frequencies, pages 122-171.
- Measuring Forces and Wear in Telephone Switching Apparatus, New Techniques for, *W. P. Mason and S. D. White*, pages 469-503.
- Mechanical Properties of Polymers at Ultrasonic Frequencies, *W. P. Mason and H. J. McSkimin*, pages 122-171.
- Menard, J. Z.*, A New Recording Medium for Transcribed Message Services, pages 530-540.
- Message Services, Transcribed
- A New Recording Medium for Transcribed Message Services, *J. Z. Menard*, pages 530-540.
- Microwaves
- Propagation Studies at Microwave Frequencies by Means of Very Short Pulses, *L. E. DeLange*, pages 91-103.
- Selective Fading of Microwaves, *A. B. Crawford and W. C. Jakes, Jr.*, pages 68-90.
- Miniaturization
- Present Status of Transistor Development, *J. A. Morton*, pages 411-442.
- Mittag-Leffler's Theorem
- Network Representation of Transcendental Impedance Functions, *M. K. Zinn*, pages 378-404.
- Mobile Radio Transmission at 150, 450, 900, and 3700 Mc., *W. R. Young, Jr.*, pages 1068-1085.
- Monfort, L. R. and Edwards, P. G.*, The Type-O Carrier System, pages 688-723.
- Montgomery, H. C.*, Electrical Noise in Semiconductors, pages 950-975.
- Morgan, S. P., Jr.*, Mathematical Theory of Laminated Transmission Lines, pages 883-949; 1121-1206.
- Morton, J. A.*, Present Status of Transistor Development, pages 411-442.
- Myers, Oscar*, Common Control Telephone Switching Systems, pages 1086-1120.

N

- Nationwide Numbering Plan, *W. H. Nunn*, pages 851-859.
- Network Representation of Transcendental Impedance Functions, *M. K. Zinn*, pages 378-404.
- Network Synthesis Using Tchebycheff Polynomial Series, *Sidney Darlington*, pages 613-665.
- Noise Theory
- Electrical Noise in Semiconductors, *H. C. Montgomery*, pages 950-975.
- Numbering Plan, Telephone
- Nationwide Numbering Plan, *W. H. Nunn*, pages 851-859.
- Nunn, W. H.*, Nationwide Numbering Plan, pages 851-859.

O

- Ohl, R. S. and Kingsbury, E. F.*, Photoelectric Properties of Ionically Bombarded Silicon, pages 802-815.
- Ohl, R. S.*, Properties of Ionic Bombarded Silicon, pages 104-121.
- Oliver, B. M.*, Efficient Coding, pages 724-750.
- Osborne, H. S. and Clark, A. B.*, Automatic Switching for Nationwide Telephone Service, pages 823-831.

P

- Photoelectric Properties of Ionically Bombarded Silicon, *E. F. Kingsbury and R. S. Ohl*, pages 802-815.

- Pilliod, J. J.*, Fundamental Plans for Toll Telephone Plant, pages 832-850.
- Plastics
Interaction of Polymers and Mechanical Waves, *W. O. Baker and J. H. Heiss*, pages 306-356.
- Polymers
Interaction of Polymers and Mechanical Waves, *W. O. Baker and J. H. Heiss*, pages 306-356.
Mechanical Properties of Polymers at Ultrasonic Frequencies, *W. P. Mason and H. J. McSkimin*, pages 122-171.
- Present Status of Transistor Development, *J. A. Morton*, pages 411-442.
- Propagation Studies at Microwave Frequencies by Means of Very Short Pulses, *O. E. DeLange*, pages 91-103.
- Properties of Ionic Bombarded Silicon, *R. S. Ohl*, pages 104-121.
- Pulse Measurements
Propagation Studies at Microwave Frequencies by Means of Very Short Pulses, *O. E. DeLange*, pages 91-103.

R

- Radio. Fading
Selective Fading of Microwaves, *A. B. Crawford and W. C. Jakes, Jr.*, pages 68-90.
- Radio Transmission
Propagation Studies at Microwave Frequencies by Means of Very Short Pulses, *DeLange*, pages 91-103.
Selective Fading of Microwaves, *A. B. Crawford and W. C. Jakes, Jr.*, pages 68-90.
- Radiotelephone Service. Mobile
Comparison of Mobile Radio Transmission at 150, 450, 900, and 3700 Mc., *W. R. Young, Jr.*, pages 1068-1085.
- Rea, W. T., Hysko, J. L. and Roberts, L. C.*, A Carrier Telegraph System for Short-Haul Applications, pages 666-687.
- Realizability Theory
Introduction to Formal Realizability Theory, *Brockway McMillan*, pages 217-279. 541-600.
- Rebound of Relay Armature
Relay Armature Rebound Analysis, *E. E. Sumner*, pages 172-200.
- Recording Medium for Transcribed Message Services, *A New, J. Z. Menard*, pages 530-540.
- Relaxation
Mechanical Properties of Polymers at Ultrasonic Frequencies, *W. P. Mason and H. J. McSkimin*, pages 122-171.
- Relay Armature Rebound Analysis, *E. E. Sumner*, pages 172-200.
- Relays
An Application of Boolean Algebra to Switching Circuit Design, *R. E. Staehler*, pages 280-305.
Important Design Factors Influencing Reliability of Relays, *J. R. Fry*, pages 976-998.
- Relays. Electromagnetic—Types AF, AG and AJ
A New General Purpose Relays for Telephone Switching Systems, *A. C. Keller*, pages 1023-1067.
- Reliability of Telephone Traffic Load Measurements by Switch Counts, *W. S. Hayward, Jr.*, pages 357-377.
- Roberts, L. C., Hysko, J. L. and Rea, W. T.*, A Carrier Telegraph System for Short-Haul Applications, pages 666-687.

S

- Sampling
The Reliability of Telephone Traffic Load Measurements by Switch Counts, *W. S. Hayward, Jr.*, pages 357-377.
- Schelkunoff, S. A.*, Generalized Telegraphist's Equations for Wave-guides, pages 784-801.
- Selective Fading of Microwaves, *A. B. Crawford and W. C. Jakes, Jr.*, pages 68-90.
- Semiconductors
Electrical Noise in Semiconductors, *H. C. Montgomery*, pages 950-975.
- Shannon-Fano Code
Efficient Coding, *B. M. Oliver*, pages 724-750.

Shipley, F. F., Automatic Toll Switching Systems, pages 860-882.

Signals, Dialing

An Experimental Electronically Controlled Automatic Switching System, *W. A. Malthaner and H. E. Vaughan*, pages 443-468.

Silicon, Electrical Properties

Properties of Ionic Bombarded Silicon, *R. S. Ohl*, pages 104-121.

Silicon, Photoelectric Properties

Photoelectric Properties of Ionically Bombarded Silicon, *E. F. Kingsbury and R. S. Ohl*, pages 802-815.

Solids

Mechanical Properties of Polymers at Ultrasonic Frequencies, *W. P. Mason and H. J. McSkimin*, pages 122-171.

Solutions

Interaction of Polymers and Mechanical Waves, *W. O. Baker and J. H. Heiss*, pages 306-366.

Mechanical Properties of Polymers at Ultrasonic Frequencies, *W. P. Mason and H. J. McSkimin*, pages 122-171.

Staehler, R. E., An Application of Boolean Algebra to Switching Circuit Design, pages 280-305.

Statistics of Television Signals, *E. R. Kretzmer*, pages 751-763.

Strains in Cable Sheaths and Other Buckled Surfaces, Principal, *I. L. Hopkins*, pages 523-529.

Sumner, E. E., Relay Armature Rebound Analysis, pages 172-200.

Switch Counts, Traffic

The Reliability of Telephone Traffic Load Measurements by Switch Counts, *W. S. Hayward, Jr.*, pages 357-377.

Switches, Crossbar

Automatic Toll Switching Systems, *F. F. Shipley*, pages 860-882.

Switching

A New General Purpose Relay for Telephone Switching Systems, *A. C. Keller*, pages 1023-1067.

An Experimental Electronically Controlled Automatic Switching System, *W. A. Malthaner and H. E. Vaughan*, pages 443-468.

Automatic Switching for Nationwide Telephone Service, *A. B. Clark and H. S. Osborne*, pages 823-831.

Automatic Toll Switching Systems, *F. F. Shipley*, pages 860-882.

Common Control Telephone Switching Systems, *Oscar Myers*, pages 1086-1120.

Fundamental Plans for Toll Telephone Plant, *J. J. Pilliod*, pages 832-850.

Nationwide Numbering Plan, *W. H. Nunn*, pages 851-859.

Transistors in Switching Circuits, *A. E. Anderson*, pages 1207-1249.

Switching Circuit Design

An Application of Boolean Algebra to Switching Circuit Design, *R. E. Staehler*, pages 280-305.

Switching Equipment

New Techniques for Measuring Forces and Wear in Telephone Switching Apparatus, *W. P. Mason and S. D. White*, pages 469-503.

T

Tchebycheff Polynomial Series

Network Synthesis Using Tchebycheff Polynomial Series, *Sidney Darlington*, pages 613-665.

Telegraph Systems, Carrier-40C1

A Carrier Telegraph System for Short-Haul Applications, *J. L. Hysko, W. T. Rea and L. C. Roberts*, pages 666-687.

Telegraphist's Equations for Waveguides, Generalized, *S. A. Schelkunoff*, pages 784-801.

Telephone Apparatus

New Techniques for Measuring Forces and Wear in Telephone Switching Apparatus, *W. P. Mason and S. D. White*, pages 469-503.

Telephone Calls

Automatic Switching for Nationwide Telephone Service, *A. B. Clark and H. S. Osborne*, pages 823-831.

- Dialing Habits of Telephone Customers, *Charles Clos and R. I. Wilkinson*, pages 32-67.
- Telephone Circuits
 An Experimental Electronically Controlled Automatic Switching System, *W. A. Malthaner and H. E. Vaughan*, pages 443-468.
- Telephone Codes
 Nationwide Numbering Plan, *W. H. Nunn*, pages 851-859.
- Telephone Service. Testing
 Dialing Habits of Telephone Customers, *Charles Clos and R. I. Wilkinson*, pages 32-67.
- Telephone Systems, Carrier—Type O
 The Type-O Carrier System, *P. G. Edwards and L. R. Montfort*, pages 688-723.
- Telephone Systems, Dial
 Common Control Telephone Switching Systems, *Oscar Myers*, pages 1086-1120.
- Telephone Systems, Toll
 Fundamental Plans for Toll Telephone Plant, *J. J. Pilliod*, pages 832-850.
- Telephone Traffic
 The Reliability of Telephone Traffic Load Measurements by Switch Counts, *W. S. Hayward, Jr.*, pages 357-377.
- Telephone Transmission
 Fundamental Plans for Toll Telephone Plant, *J. J. Pilliod*, pages 832-850.
- Teletypewriters
 A Carrier Telegraph System for Short-Haul Applications, *J. L. Hysko, W. T. Rea and L. C. Roberts*, pages 666-687.
- Television Signals
 Experiments with Linear Prediction in Television, *C. W. Harrison*, pages 764-783.
 Statistics of Television Signals, *E. R. Kretzmer*, pages 751-763.
- Temperature Measurements
 Properties of Ionic Bombarded Silicon, *R. S. Ohl*, pages 104-121.
- Thirtieth Anniversary (of The Bell System Technical Journal), page 611.
- Toll Traffic
 Automatic Toll Switching Systems, *F. F. Shipley*, pages 860-882.
 Fundamental Plans for Toll Telephone Plant, *J. J. Pilliod*, pages 832-850.
- Transistors
 Electrical Noise in Semiconductors, *H. C. Montgomery*, pages 950-975.
- Transistors, Types M1689, M1698, M1729, M1734, M1752
 Present Status of Transistor Development, *J. A. Morlon*, pages 411-442.
- Transistors in Switching Circuits, *A. E. Anderson*, pages 1207-1249.
- Translators
 Automatic Toll Switching Systems, *F. F. Shipley*, pages 860-882.
- Transmission Lines
 Network Representation of Transcendental Impedance Functions, *M. K. Zinn*, pages 378-404.
- Transmission Lines, Clogston 1 and 2
 Mathematical Theory of Laminated Transmission Lines, *S. P. Morgan, Jr.*, pages 883-949; 1121-1206.
- Transmission Measurements
 Comparison of Mobile Radio Transmission at 150, 450, 900, and 3700 Mc., *W. R. Young, Jr.*, pages 1068-1085.
- Trunking
 Dialing Habits of Telephone Customers, *Charles Clos and R. I. Wilkinson*, pages 32-67.
- Type-O Carrier System, *P. G. Edwards and L. R. Montfort*, pages 688-723.

V

- Vaughan, H. E. and Malthaner, W. A.*, An Experimental Electronically Controlled Automatic Switching System, pages 443-468.
- Viscosity
 Interaction of Polymers and Mechanical Waves, *W. O. Baker and J. H. Heiss*, pages 306-356.
 Mechanical Properties of Polymers at Ultrasonic Frequencies, *W. P. Mason and H. J. McSkimin*, pages 122-171.

W

Waveguides

Generalized Telegraphist's Equations for Waveguides, *S. A. Schelkunoff*, pages 784-801.

Wear Studies of Telephone Apparatus

New Techniques for Measuring Forces and Wear in Telephone Switching Apparatus, *W. P. Mason and S. D. White*, pages 469-503.

White, S. D. and Mason, W. P., New Techniques for Measuring Forces and Wear in Telephone Switching Apparatus, pages 469-503.

Wilhelm, H. T., Impedance Bridges for the Megacycle Range, pages 999-1012.

Wilkinson, R. I. and Clos, Charles, Dialing Habits of Telephone Customers, pages 32-67.

Wiring and Wrapping Tools

A New General Purpose Relay for Telephone Switching Systems, *A. C. Keller*, pages 1023-1067.

Y

Young, W. R., Jr., Comparison of Mobile Radio Transmission at 150, 450, 900, and 3700 Mc., pages 1068-1085.

Z

Zinn, M. K., Network Representation of Transcendental Impedance Functions, pages 378-404.



